Published in final edited form as:

Biom J. 2016 July ; 58(4): 974–992. doi:10.1002/bimj.201500171.

Time-dependent classification accuracy curve under marker-dependent sampling

Zhaoyin Zhu¹, Xiaofei Wang^{*,2}, Paramita Saha-Chaudhuri³, Andrzej S. Kosinski², and Stephen L. George²

¹ Division of Biostatistics, New York University School of Medicine, New York, NY 10016, USA

² Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705, USA

³ Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC H3A 1A2, Canada

Abstract

Evaluating the classification accuracy of a candidate biomarker signaling the onset of disease or disease status is essential for medical decision making. A good biomarker would accurately identify the patients who are likely to progress or die at a particular time in the future or who are in urgent need for active treatments. To assess the performance of a candidate biomarker, the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are commonly used. In many cases, the standard simple random sampling (SRS) design used for biomarker validation studies is costly and inefficient. In order to improve the efficiency and reduce the cost of biomarker validation, marker-dependent sampling (MDS) may be used. In a MDS design, the selection of patients to assess true survival time is dependent on the result of a biomarker assay. In this article, we introduce a nonparametric estimator for time-dependent AUC under a MDS design. The consistency and the asymptotic normality of the proposed estimator is established. Simulation shows the unbiasedness of the proposed estimator and a significant efficiency gain of the MDS design over the SRS design.

Keywords

Biomarker; Classification accuracy; Marker-dependent sampling; Smoothing; Time-dependent AUC; Time-to-event data

1 Introduction

In the era of personalized medicine, it is critically important to determine the prognostic value of candidate biomarkers. A biomarker with good prognostic value can help identify patients at high risk of early recurrence or progression and treat these patients with

*Corresponding author: xiaofei.wang@duke.edu, Phone: +1-919-6815406.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

Conflict of interest

The authors have declared no conflict of interest.

aggressive treatments, while patients at low risk of early recurrence or progression will be put on less aggressive treatments to avoid unnecessary adverse effects caused by such treatments. Evaluation of the accuracy of a candidate biomarker that can predict the onset of disease or progression to death is an essential goal of studies of biomarkers. The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) are widely used to assess the performance of a diagnostic test or a biomarker. In addition to the presence or absence of the disease, the timing of disease transition is often of interest. In this article, we are interested in the problem of utilizing a biomarker to classify patients with respect to a time-to-event outcome, such as progression-free survival. Progression-free survival is a continuous measure of the time from diagnosis to disease progression. Time-dependent accuracy measures generalize the notion of classification accuracy for binary classification of disease status to a time-to-event endpoint. The time-dependent ROC and AUC were proposed (Heagerty and Zheng, 2005; Saha-Chaudhuri and Heagerty, 2013) to characterize how well a candidate biomarker performs in predicting which patients will experience a transition or who will remain event-free over time. The method avoids loss of information and potential bias caused by dichotomization of survival time into a binary event status, that is being event-free less or greater than a specified time point.

In the simple random sampling (SRS) design, subjects are randomly sampled from the population regardless of their biomarker values. Under a SRS design, to evaluate the classification accuracy of candidate biomarkers for survival time, patients must be followed for a long time that can be costly and time consuming. This article considers an alternative study design called the marker-dependent sampling (MDS) design. The MDS design allows investigators to divide patients into subgroups based on their biomarker values, and oversample the patient subgroups that carry more information while undersample patient subgroups with less information.

The idea underlying the MDS design comes from a long tradition in biomedical research utilizing target sampling to improve study efficiency. For example, in assessing the relationship between risk and exposure, case-control sampling for a binary outcome (e.g. Breslow et al., 1980) and outcome-dependent sampling for a continuous outcome (e.g. Weaver and Zhou, 2005; Wang and Zhou, 2010) have often been used to oversample subjects with more information and increase study efficiency with fixed cost. The MDS design can be considered as an extension of the idea of target sampling from the assessment of risk-exposure association to the assessment of biomarker-disease classification. In this article, we consider a specific MDS design that consists of a SRS component and a MDS component. The SRS component is a sample of subjects randomly selected from the target population while the MDS component consists of multiple samples of subjects selected from strata defined by the intervals of the biomarker value based on the goal of the study. If the investigators are particularly interested in the biomarker performance in a certain region of the biomarker values, subjects within this region of interest may be oversampled. If the investigators are interested in the biomarker performance in the entire range of the biomarker values, subjects in low, middle, and high range may be sampled with balanced allocation to maximize the estimation efficiency. As demonstrated in Wang et al. (2012, 2013), the MDS design requires less patients for enrollment and follow-up and can improve the efficiency in assessing the overall performance of candidate biomarkers. In biomedical

literature, similar ideas have been successfully used to improve efficiency of evaluating the accuracy of a biomarker in classifying disease status (e.g. Morra et al., 2007; Strauss et al., 2010; Williams et al., 2014; Selen et al., 2014; Ding et al., 2014; Schildcrout et al., 2015).

We develop a nonparametric estimator of time-dependent AUC under the MDS design, which has fewer assumptions than parametric or semiparametric estimators. The proposed estimator estimates AUC as time-dependent measures for classification accuracy at each time point of the follow-up period without losing information due to dichotomization and introducing potential bias due to ignoring censored patients. Bamber (1975) and Hanley and McNeil (1982) developed the nonparametric AUC estimator for binary disease status. Heagerty and Zheng (2005) defined time-dependent AUC and proposed a semiparametric estimator under a proportional hazard assumption. Saha-Chaudhuri and Heagerty (2013) studied a nonparametric estimator using weight mean rank of neighboring patients at risk to estimate time-dependent AUC and applied it to compare different biomarkers. Wang et al. (2012, 2013) discussed the nonparametric estimation of ROC, AUC, and partial AUC under the MDS design when disease status is a binary variable.

The rest of the article is organized as follows. In Section 2, we briefly review the definition of time-dependent AUC and the nonparametric estimator under the SRS design. We propose the non-parametric estimation of time-dependent AUC under marker-dependent sampling and describe its asymptotic properties in Section 3. The consistency and efficiency gain of the proposed AUC estimator and the performance of the proposed bootstrap variance estimator in finite samples is evaluated through simulation study in Section 4. A discussion is given in Section 5. Proofs of the asymptotic properties are provided in the Appendix.

2 Nonparametric estimation of time-dependent AUC

Let n denote the total size of the cohort study. Let T_i^* denote the true survival time and C_i the independent censoring time for patient i . We assume T_i^* and C_i are independent and identically distributed, respectively. We observe the follow-up time $T_i = \min \{T_i^*, C_i\}$ and the censoring indicator $\Delta_i = \mathbb{I} \{T_i^* \leq C_i\}$. Δ_i equals 0 if a patient is censored and 1 otherwise. Let $R_i(t) = \mathbb{I} \{T_i \geq t\}$ denote the at-risk indicator. Cases and controls any time t are defined over the at-risk set. $\mathcal{R}_t = \{i: R_i(t) = 1\}$. Let $D_i^1(t) = \mathbb{I} \{R_i(t) = 1, T_i^* = 1\}$ be the indicator for cases and $\mathcal{R}_t^1 = \{i: D_i^1(t) = 1\}$ be the set of patients who have event at t . Let $D_i^0(t) = \mathbb{I} \{R_i(t) = 1, T_i^* > t\}$ be the indicator for controls and $\mathcal{R}_t^0 = \{i: D_i^0(t) = 1\}$ be the set of patients who do not have event at t . Let M_i be independently continuous measures for the candidate biomarker, where a higher value indicates more severe disease and a shorter survival time. Etzioni et al. (1999) and Slate and Turnbull (2000) adopted a definition of time-dependent sensitivity and specificity at the threshold c :

$$\begin{aligned} \text{Sensitivity}(c, t): TP_t(c) &= P(M_i > c | T_i^* = t) \\ \text{Specificity}(c, t): FP_t(c) &= P(M_i \leq c | T_i^* > t) \end{aligned}$$

Heagerty and Zheng (2005) defined the time-dependent ROC and AUC:

$$ROC_t(p) = TP_t \left\{ [FP_t]^{-1}(p) \right\}$$

$$AUC(t) = \int_0^1 TP_t \left\{ [FP_t]^{-1}(p) \right\} dp = Pr \left(M_i > M_j | T_i^* = t, T_j^* > t \right)$$

For time-dependent AUC, each time point can be considered as a case of simple random sample, the nonparametric estimator of time-dependent AUC at time t is given by:

$$A(t) = \left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t) D_j^0(t) I \{M_i > M_j\} \right] / \left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t) D_j^0(t) \right]$$

However, commonly only a small number of patients would experience an event at any given time t . Therefore to obtain desirable statistical properties, Saha-Chaudhuri and Heagerty (2013) developed a nonparametric smoothing estimator of time-dependent AUC:

$$\hat{AUC}(t) = \frac{1}{|\mathcal{N}_t(h_n)|} \sum_{t_i \in \mathcal{N}_t(h_n)} A(t_i),$$

where $\mathcal{N}_t(h_n) = \{t_i : |t - t_i| < h_n\}$ denotes a neighborhood around t .

3 Nonparametric estimation of time-dependent AUC under MDS

3.1 MDS design and notation

We consider a MDS design that consists of two parts: one is a simple random component and the other is a marker-dependent component. Assume the biomarker can be clustered into K mutually exclusive intervals $C_k = (a_{k-1}, a_k]$, $k = 1, 2, \dots, K$ and $a_0 = -\infty$, $a_K = \infty$. Let n_0 denote the number of patients of the simple random component. For the patients of the marker-dependent component, let n_k , $k = 1, 2, \dots, K$ denote the number of patients whose biomarker value falls into the interval C_k . The total size of the cohort is $n = \sum_{s=0}^K n_s$. We let $f(M, T)$ be the joint density function of biomarker M and survival time T . Let $p_{si} = f(M_{si}, T_{si})$, $s = 0, 1, \dots, K$, $i = 1, 2, \dots, n_s$, and θ_k be the probability that biomarker M falls into the interval C_k , $\theta_k = Pr(M \in C_k)$, $k = 1, 2, \dots, K$.

3.2 Estimation of MDS distribution

We observe the marker-dependent component $\{M_{ki}, T_{ki}\}$, $k = 1, 2, \dots, K$, $i = 1, 2, \dots, n_s$, conditional on whether M belongs to the interval C_k as well as the simple random component $\{M_{0i}, T_{0i}\}$, where $i = 1, 2, \dots, n_0$. Let $p = \{p_{si}\}$, $s = 0, 1, \dots, K$, $i = 1, 2, \dots, n_s$, and $\theta = \{\theta_k\}$, $k = 1, 2, \dots, K$. Thus the empirical likelihood of $\{M_{ki}, T_{ki}\}$ is:

$$L(p, \theta) = \left[\prod_{i=1}^{n_0} f(M_{0i}, T_{0i}) \right] \left[\prod_{k=1}^K \prod_{j=1}^{n_k} f(M_{kj}, T_{kj} | M_{kj} \in C_k) \right] = \left[\prod_{s=1}^K \prod_{i=1}^{n_s} p_{si} \right] \left[\prod_{k=1}^K \theta_k^{-n_k} \right] \quad (1)$$

We search $\{p, \theta\}$ that maximizes the equation (1) with the constraints:

$$\left\{ p_{si} \geq 0, \sum_{s=0}^K \sum_{i=1}^{n_s} p_{si} = 1, \sum_{s=0}^K \sum_{i=1}^{n_s} p_{si} (\mathbb{I}(M_{si} \in C_k) - \theta_k) = 0 \right\}$$

Wang and Zhou (2006) solved this problem by adding Lagrange multipliers λ_k and μ .

$$H(p, \theta, \lambda, \mu) = \sum_{s=1}^K \sum_{i=1}^{n_s} \log p_{si} - \sum_{k=1}^K (n_k \log \theta_k) + \sum_{k=1}^K \lambda_k \sum_{s=0}^K \sum_{i=1}^{n_s} p_{si} (\mathbb{I}(M_{si} \in C_k) - \theta_k) + \mu \left(1 - \sum_{s=0}^K \sum_{i=1}^{n_s} p_{si} \right) \quad (2)$$

Taking derivatives with respect to $\{p, \theta, \lambda, \mu\}$ respectively, we obtain the estimator of p_{si} :

$$\hat{p}_{si} = \frac{1}{n_0 + \sum_{k=1}^K \frac{n_k}{\theta_k} \mathbb{I}(M_{si} \in C_k)} \quad (3)$$

Plugging (3) into (2), we are able to obtain the estimator of θ_k by solving $\frac{\partial l_p(\theta)}{\partial \theta} = 0$, where

$$l_p(\theta) = - \sum_{s=0}^K \sum_{i=1}^{n_s} \log \left[n_0 + \sum_{k=1}^K \frac{n_k}{\theta_k} \mathbb{I}(M_{si} \in C_k) \right] - \sum_{k=1}^K (n_k \log \theta_k)$$

We can order patients by their second subscript $i, i = 1, 2, \dots, n_s$, within each interval and order the intervals by the first subscript $s, s = 0, 1, \dots, K$. Since $\sum_{s=1}^K \sum_{i=1}^{n_s} n_{si} = n$, we are able to replace the double subscripts si by a single subscript $i, i = 1, 2, \dots, n$ for notation simplicity. The sequence $\{p_{si}\}$ is now written as $\{p_i\}, i = 1, 2, \dots, n$ ordered by corresponding patient id.

3.3 Nonparametric estimation of AUC(t) and asymptotic properties

Under the SRS design, Saha-Chaudhuri and Heagerty (2013) showed that the nearest neighbor estimator

$$\hat{A}^{SRS}(t) = \frac{1}{|\mathcal{N}_t(h_n)|} \sum_{t_i \in \mathcal{N}_t(h_n)} A^{SRS}(t_i)$$

where

$$A^{SRS}(t_i) = \left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t_i) D_j^0(t_i) \mathbb{I}\{M_i > M_j\} \right] / \left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t_i) D_j^0(t_i) \right]$$

is equivalent to the Nadaraya–Watson estimator, and the bias is negligible when n is large. However, under the MDS design, $A(t)$ will yield bias in estimating the AUC at target time t due to the nonrandom selection of patients. In order to correct the bias, we utilize the estimates $\hat{p}_{sj}$ and develop the nonparametric estimator under the MDS design:

$$\hat{A}^{MDS}(t) = \frac{1}{|\mathcal{N}_t(h_n)|} \sum_{t_i \in \mathcal{N}_t(h_n)} A^{MDS}(t_i)$$

where

$$A^{MDS}(t_i) = \left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t_i) D_j^0(t_i) \hat{p}_i \hat{p}_j \mathbb{I}\{M_i > M_j\} \right] / \left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t_i) D_j^0(t_i) \hat{p}_i \hat{p}_j \right]$$

At a fixed target time t , our proposed estimator $\hat{A}^{MDS}(t)$ can be rewritten as a weighted U -statistics of order $k = 2$. By the theorem from Lee (1990), we are able to prove that the proposed estimator converges in distribution to standard normal distribution:

$$V_n^{-1/2} \left[\hat{A}^{MDS}(t) - AUC(t) - b_n(t) \right] \rightarrow_d N(0, 1)$$

where $V_n = \text{var}[\hat{A}^{MDS}(t)]$ and $b_n(t)$ denotes the bias due to smoothing, which can be asymptotically negligible relative to $V_n^{-1/2}$. See the Appendix for details. To estimate the standard error of our proposed estimator, we use a stratified bootstrapping algorithm (Efron, 1979) to reach a quick approximation. The corresponding point-wise confidence intervals are obtained by the 2.5 percentile and the 97.5 percentile of the estimators from bootstrap method. More details are given in the simulation section.

3.4 Selection of bandwidth

Our proposed estimator $\hat{A}^{MDS}$ is a nearest neighbor estimator with a kernel smoothing technique. Selection of optimal bandwidth is essential for good performance of the estimator. It is expected that a local bandwidth specific for time t offers better performance for the MDS design than a global bandwidth as in Saha-Chaudhuri and Heagerty (2013). Instead of selecting a global bandwidth that minimizes the mean integrated squared error (MISE), we choose a local bandwidth $h_n^L(t)$ that minimizes the exact mean square error (MSE) of $\hat{A}^{MDS}(t)$ at each time point:

$$h_n^L(t) = \argmin_{h_n > 0} \text{MSE} \left[\hat{A}^{MDS}(t) \right]$$

In a specific data analysis, cross-validation based on a grid search can be applied to obtain the asymptotically optimal bandwidth. Staniswalis (1989) proposed a consistent estimate of exact MSE for local bandwidth selection:

$$M\hat{S}E\left[\hat{A}^{MDS}(t)\right]=\left[\tilde{A}^{MDS}(t)-\tilde{A}(t)\right]^2,$$

where $\tilde{A}(t)$ is a kernel estimate that uses the global optimal bandwidth h_n^G that is selected by leave-one-out cross-validation and $\tilde{A}^{MDS}(t)=\frac{1}{|\mathcal{N}_t(h_n)|}\sum_{t_i\in\mathcal{N}_t(h_n)}\tilde{A}(t_i)$. Then the optimal local bandwidth at target point t is

$$h_n^L(t)=\operatorname{argmin}_{h_n>0}M\hat{S}E\left[\hat{A}^{MDS}(t)\right].$$

4 Simulation study

4.1 Simulation design

To evaluate the performance of the estimator $A^{MDS}(t)$ under the MDS design, we conducted a series of simulation studies. In this section, we show the unbiasedness of the proposed estimator and demonstrate the efficiency gain of the MDS design over a SRS design. We use a stratified bootstrapping method to obtain the standard error of the proposed estimator. Additionally, the simulation confirms that our estimator performs well under different scenarios, encompassing different allocation of the MDS components and the SRS component and different censoring percentages.

Let T denote the log event time and M denotes the value of biomarker. For the purpose of simulation studies, we assume the joint distribution of (T, M) follows a bivariate normal distribution $N(0, 0, 1, 1, -0.7)$ where means equal 0, variances equal 1 and the correlation between T and M equals -0.7 . We suppose 20% of patients in the cohort are censored and an independent log censoring time that follows a standard normal distribution is generated. The total size of study cohort is $n=360$, in which 180 subjects come from the SRS component and the other 180 subjects come from the MDS component. The intervals of the MDS component are generated as following: $C_1=(-\infty, a_1)$, $C_2=[a_1, a_2]$, $C_3=(a_2, \infty)$, where $a_1=\mu_M-\sigma_M$ and $a_2=\mu_M+\sigma_M$. The quantities μ_M and σ_M denote the mean and standard derivation of the biomarker. Compared to a SRS design, we oversample the patients from two tails, we assume the number of patients in C_1 , C_2 , C_3 are (60, 60, 60). All the results are based on 1000 independent simulations and for each simulation 2000 bootstrap resamples are carried out to estimate standard error and confidence interval.

4.2 Unbiasedness of proposed estimator $\hat{A}^{MDS}(t)$

Table 1 shows that under the MDS design, the SRS estimator $\hat{A}^{SRS}(t)$ that does not consider the MDS data structure would yield bias, while the proposed MDS estimator $\hat{A}^{MDS}(t)$ performs much better with almost no bias. For example, at $\log(t)=0$, the bias of $\hat{A}^{SRS}(t)$ is as big as 5.6% and the bias of $\hat{A}^{MDS}(t)$ has been reduced to $<0.1\%$.

4.3 Estimation of standard error under MDS

Table 2 shows the target log time ($\log(t)$), the true AUC(t) ($AUC(t)$) and the estimates $\hat{A}^{MDS}$, the bias (Bias%), true or Monte Carlo standard error (SE) of our proposed estimator under the MDS design. To obtain standard error of $\hat{A}^{MDS}$, the nonparametric bootstrap method based on resampling with replacement within each SRS or MDS stratum is used. We also listed the bootstrap estimates (Bootstrap $\hat{A}^{MDS}$, bias (Bootstrap Bias), standard error (Bootstrap SE), and empirical 95% coverage percentage (95% CP).

As seen in Table 2, the bootstrap methods perform well in estimating the standard error of the proposed estimator and the bootstrap SE is very close to the Monte Carlo SE. We obtained approximately 95% coverage probability in each target time point except at the very early time that may be due to the inadequate number of events.

4.4 Efficiency comparison of MDS design with SRS design

One of the benefits of the MDS design over the standard SRS design is that the MDS design is able to improve the efficiency in assessing the performance of candidate biomarker with the same sample size. In this simulation study, we generated a dataset for the SRS design with sample size $n = 360$ as well as a dataset for the MDS design consisting of a 180 SRS component and a 180 MDS component. Table 3 shows the performance of the SRS estimator $\hat{A}^{SRS}(t)$ under the SRS design and that of the MDS estimator of $\hat{A}^{MDS}(t)$ under marker-dependent sampling. The results in Table 3 confirm that the MDS design is able to increase the study efficiency and reduce the cost compared to the SRS design. The results are understandable since the MDS design allows investigators to select more patients from those in the intervals of biomarker value that contain more information for statistical estimation at the design phase of the study.

4.5 Performance of $\hat{A}^{MDS}(t)$ under other scenarios

To demonstrate the robustness of the proposed estimator $\hat{A}^{MDS}(t)$ and to search for the optimal parameters for the MDS design, we carried out additional simulations under different scenarios. Table 4 lists the bias and RMSE for the proposed estimator, standard SRS estimator under different scenarios. This table shows that the estimator $\hat{A}^{MDS}$ under the MDS data is relatively unbiased compared to $\hat{A}^{MDS}$ and more efficient than $\hat{A}^{SRS}$ under the SRS design for all the listed scenarios.

When inspecting the RMSE column of $\hat{A}^{MDS}$, we can find that enlarging the allocation of the MDS component in the MDS design does not result in an improvement in efficiency. However, if we oversample the patients with higher values of biomarkers, significant efficiency gain at early time point is detected accompanying with the relatively low efficiency at late time point. This result is understandable since higher values represent short-survival time and the number of patients at risk decreases rapidly. Additionally, a high percentage of censoring will increase the RMSE of the estimated AUC(t) because more information is lost.

5 Data example

The primary motivation of adopting a MDS design over the SRS design is the potential of an improved efficiency associated with the MDS design in estimating time-dependent $AUC(t)$. In practice, data with MDS structure can also arise from pooled analysis of individual patient data assembled from multiple sources. In this section, we illustrate the proposed $AUC(t)$ estimator with the analysis of prognostic value of COX-2 using the pooled data from two randomized clinical trials conducted by the Cancer and Leukemia Group B (CALGB).

COX-2 is a protein that is commonly overexpressed in the serum of lung cancer patients. COX-2 level is measured on a continuous scale ranging from 0 to 10. Higher COX-2 level is believed to be associated with worse survival. CALGB 30203 (Edelman et al., 2007) is a phase II trial designed to evaluate the effect of Celecoxib, a COX-2 inhibitor, in treating advanced nonsmall cell lung cancer. Eighty (80) patients were enrolled and their COX-2 levels were determined. These patients were randomized to receive either Celecoxib (a COX-2 inhibitor) or Placebo. As a follow-up trial, CALGB 30801 (Edelman et al., 2014), was conducted to further evaluate the effect of Celecoxib in prolonging progression-free survival (PFS) relative to Placebo in the same population. Since the effect of Celecoxib is of more interest for the patients with moderate or higher COX-2 levels, a total 316 patients with $COX-2 \geq 2$ were randomized and received protocol treatments. Median PFS and its 95% confidence interval (CI) in months by treatment arm and interval of COX-2 score are given in Table 5.

We are interested in validating the prognostic value of COX-2 values using the data from both trials. To assess the prognostic value of COX-2, we should use all patients treated on the Placebo arm of both trials. Given that there is no statistically difference between two arms in CALGB 30801 (log rank test, $P = 0.81$), we include all CALGB 30801 patients in the analysis. For CALGB 30203, Edelman et al. (2007) reported a significant benefit of Celecoxib in treating patients with $COX-2 \leq 4$ (log rank test, $P = 0.002$), but again no statistically significant difference in PFS between two arms when all patients are considered (log rank test, $P = 0.13$). For these reasons, we include all patients from the two trials in the illustration of the proposed $AUC(t)$ estimator. Specifically, CALGB 30203 patients are considered as a simple random sample from the target patient population, as all screened patients were included regardless of their COX-2 levels; CALGB 30801 patients are considered as a marker-dependent sample (MDS), as only those with $COX-2 \geq 2$ were randomized and followed for progression-free survival (PFS). The pooled data of the two trials have the same structure as the MDS design and consistent estimation for $AUC(t)$ can be obtained using the proposed estimator. All the patients with valid data can be classified into three categories by COX-2: negative ($COX-2 < 2$), moderate ($2 \leq COX-2 < 4$), and positive ($COX-2 \geq 4$). The simple random component includes n_0 82 patients with $n_{0,1} = 82$ patients with $n_{0,1} = 32$ negative, $n_{0,2} = 18$ moderate and $n_{0,3} = 32$ positive. The marker-dependent component includes 305 patients and its corresponding strata are $n_1 = 0$ negative, $n_2 = 83$ moderate, and $n_3 = 222$ positive. Figure 1 shows the estimated $AUC(t)$ at each target time point by the standard method ignoring the MDS data structure (red line) and our proposed method (black line). There is a difference between the two curves for the proposed estimator $\hat{A}^{MDS}$ and the standard estimator $\hat{A}^{SRS}$. For example, at time point $t = 11$, $\hat{A}^{MDS}$

equals 0.510 with corresponding confidence interval (0.441, 0.598) and the biased estimate $\hat{A}^{SRS}$ equals 0.473.

6 Discussion

In this article, we introduce an efficient study design for validating the classification accuracy of a candidate biomarker for survival time. The marker-dependent sampling (MDS) design consists of two sampling components, a simple random sample and multiple marker-dependent samples. The MDS study design is motivated by the limitation of the standard simple random sampling (SRS) design that has low efficiency when the biomarkers in certain ranges of interest occur much less frequently in the population. The MDS design, utilizing the value of a biomarker, oversamples those patients with more information and undersamples those patients with less or redundant information. Accordingly, the MDS design can reduce the number of patients to be followed for long-term survival and reduce the overall study cost. Under the MDS design, we develop a nonparametric estimator for time-dependent AUC to evaluate the accuracy of a biomarker. We demonstrate the asymptotic properties of our proposed estimator including unbiasedness and normality. Simulations confirm that our proposed estimator successfully corrects the bias under the MDS design by employing the empirical likelihood estimates of the MDS distribution. Simulations also show a significant efficiency gain of the MDS design relative to the SRS design. By adopting a bootstrap method, we are able to estimate the variance of the estimator and obtain the corresponding point-wise 95% confidence interval with correct nominal coverage probability.

We extend the WMR estimator of Saha-Chaudhuri and Heagerty (2013) to the MDS design. The main advantage of using the WMR estimator as the basis of the proposed estimator is due to the ease and directness of computing the WMR estimator and the fact that the estimated empirical probability mass works nicely with the WMR estimator. As $AUC(t)$ is a functional of the conditional survival function, an alternative estimator for $AUC(t)$ based on the conditional survival function could be developed. However, given the conditioning event of $T = t$ for the case subject, a two-stage bandwidth selection would be required, first for estimating the conditional survival function, and then for estimating the AUC. The computation of the alternative approach would be challenging.

To utilize the MDS design, it is important to know how to conduct target sampling on the patients with more information for AUC estimation. The variance estimator of the time-dependent $AUC(t)$ is a complicated function of the number of events and the number of subjects at risk at time t and the joint distribution of survival time and the biomarker. Thus, methods for maximizing estimation precision by choosing different design parameters require careful investigation. Candidate biomarkers often follow a bell-shaped distribution with a majority of subjects falling in the middle and few subjects with low or high biomarker values. In such cases, the MDS design that oversamples extreme low/high biomarker values and undersamples midrange biomarker values will increase the representation of subjects with low/high biomarker values and accordingly enhance the precision of AUC estimation. Through simulation, we were able to obtain some heuristic answers to this important question and some results can be found in the Simulation section. In particular, we found

that dividing subjects into three groups based on biomarker values (e.g. high, moderate, or low), and evenly distributing subjects across these groups, will achieve close to optimal efficiency when estimating $AUC(t)$ using the proposed estimator. A similar observation that optimal efficiency gain for AUC occurs when biomarkers are evenly distributed across different intervals of the biomarker value was reported by Wang et al. (2012, 2013) for the problem of binary disease status. Our simulation also suggests that when the biomarker value and survival time are negatively correlated and the performance of $AUC(t)$ at early time is of particular interest, oversampling the subjects with high biomarker values will significantly increase efficiency. When a different accuracy measure, such as partial AUC, is of interest, target sampling on the subjects whose biomarker values falling into the suitable region of FPRs will effectively increase the efficiency of estimating partial AUC. More theoretical or numerical research is needed to determine the optimal parameters for the MDS design, especially considering that the MDS design may add additional cost and time in screening patients. Future research is needed to evaluate the potential loss of information resulting from ignoring the known biomarker values of unselected patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

The authors would like to acknowledge the support from the following research grants for this project: P01-CA142538 (XFW) and Duke B&B Methods Grant (X.F.W., P.S.C., A.S.K.). The authors also want to thank the associate editor and one referee for their valuable comments that improved the clarity of the article.

Appendix

Asymptotic normality of proposed estimator

$$\hat{A}^{MDS}(t) = \frac{1}{|N_t|} (h_n) \mid \sum_{t_i \in \mathcal{N}_t(h_n)} A^{MDS}(t_i)$$

where

$$A^{MDS}(t_i) = \frac{\left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t_i) D_j^0(t_i) \hat{p}_i \hat{p}_j \mathbb{I}\{M_i > M_j\} \right]}{\left[\sum_{i=1}^n \sum_{j=1}^n D_i^1(t_i) D_j^0(t_i) \hat{p}_i \hat{p}_j \right]}$$

$$\hat{p}_i = \frac{1}{n_0 + \sum_{k=1}^K \frac{n_k}{\hat{\theta}_k} (M_i \in C_k)}$$

We assume that no two subjects fail at the same time and no subjects are censored. We denote the ordered observed event times as $t_{(1)} < t_{(2)} < \dots < t_{(m)}$. Now we focus on the target time t . We restrict our attention to a neighborhood around t : $N(h_n(t)) = t_j : |t - t_j| < h_n(t)$ and $|N(h_n(t))| = m_t$. As $n \rightarrow \infty$, the bandwidth $h_n(t) = b(t) * n^{-0.2}$ decreases and the total sample size increases. However, the rate of decrease of bandwidth is smaller than the rate of

increase of sample size, $nh_n(t) \rightarrow \infty$, thus the number of subjects within the neighborhood $m_t \rightarrow \infty$ as $n \rightarrow \infty$.

Let $L_t = |t_j: t_j < t - h_n(t)|$ denote the number of failures that are observed before the start of the neighborhood of interest. The indices of the observed failure times within the neighborhood are then $L_t + 1, L_t + 2, \dots, L_t + m_t$ and that $t_{(L_t+1)} < t_{(L_t+2)} < \dots < t_{(L_t+m_t)}$ are the corresponding unique event times within $N(h_n(t))$. We assume that all event times are ordered such that for the i -th subject with event time T_i and biomarker M_i this individual corresponds to the ordered time $T_i = t_{(i)}$.

$$\begin{aligned} A^{\hat{MDS}}(t) &= \frac{1}{|N_t(h_n(t))|} \sum_{t_i \in N_t(h_n(t))} A(t_i) = \\ &= \frac{1}{m_t} \sum_{i=L_t+1}^{L_t+m_t} A(t_i) = \\ &= \frac{1}{m_t} \sum_{i=L_t+1}^{L_t+m_t} \frac{1}{\sum_{l=1}^n I(T_l > t_{(i)})} \sum_{j=i+1}^n I(M_i > M_j) I(T_i < T_j) \hat{p}_j = \\ &= \frac{1}{m_t} \sum_{i=L_t+1}^{L_t+m_t} \frac{1}{\sum_{l=1}^n I(T_l > t_{(i)}) \hat{p}_i} \sum_{j=i+1}^n I(M_i > M_j) I(T_i < T_j) \hat{p}_i \hat{p}_j = \\ &= \frac{1}{m_t} \sum_{i=L_t+1}^{L_t+m_t} \frac{1}{n_{t(i)}} \sum_{j=i+1}^n I(M_i > M_j) I(T_i < T_j) \hat{p}_i \hat{p}_j = \\ &= \sum_{i=L_t+1}^n \sum_{\substack{j=L_t+1 \\ j \neq i}}^n \frac{1}{2 \times m_t \times n_{t(i)}} I(M_i > M_j) I(T_i < T_j) \hat{p}_i \hat{p}_j, \end{aligned}$$

where $n_{t(i)} = \sum_{l=1}^n I(T_l > \min(t_{(i)}, t_{(j)})) \hat{p}_i \hat{p}_{\min(i, j)}$

For two subjects i and j with $i, j > L_t + m_t$, we cannot order the failure times. Therefore, we assume that $I(T_i < T_j) = 0$ and $I(T_i > T_j) = 0$.

We note that $A^{\hat{MDS}}(t)$ is a linear transformation of weighted U -statistic: $A^{\hat{MDS}}(t) = 1/2(U + 1)$ defined as

$$U = \sum_{i=L_t+1}^n \sum_{\substack{j=L_t+1 \\ j \neq i}}^n w_{ij} \operatorname{sgn}(j-i) \operatorname{sgn}(R_i - R_j),$$

where $\operatorname{sgn}(x) = 1$ if $x > 0$ and $\operatorname{sgn}(x) = -1$ if $x < 0$, R_j is the rank of the biomarker corresponding to the j -th ordered observed failure time and

$$w_{ij} = \frac{\hat{p}_i \hat{p}_j}{2 \times m_t \times n_{t(i)}} = \frac{\hat{p}_{\max(i, j)}}{2 \times m_t \times \sum_{l=1}^n I(T_l > \min(t_{(i)}, t_{(j)})) \hat{p}_l}.$$

Wang et al. (2012, 2013) provided the convergence and asymptotic properties of $\hat{\theta}$ and $\hat{p}_i$. Here, we state the following theorem from Lee (1990) (Theorem 1 page 153).

Let $W_n = \sum_{(n,k)} w(S) \psi(S)$ be a weighted U -statistic of order $k = 2$ and define

$$w_{i,n} = \sum_{S: i \in S} w(S),$$

$$W_n^* = \sum_{i=1}^n w_{i,n} \psi_i(X_i).$$

Suppose that the following conditions hold when $n \rightarrow \infty$:

(i)

$$\max_{1 \leq i \leq n} \frac{|w_{i,n}|^2}{\sum_{i=1}^n w_{i,n}^2} \rightarrow 0$$

(ii)

$$\frac{\sum_{(n,2)w^2(S)} w_{i,n}^2}{\sum_{i=1}^n w_{i,n}^2} \rightarrow 0$$

(iii) $E|\psi_1(X_1)|^{2+\delta} < \infty$ for some $\delta > 0$

Now we begin to verify these three conditions with our proposed estimator $A^{\hat{MDS}}(t)$

$$w_i = \sum_j w_{ij} = \frac{1}{2m_t} \sum_j \frac{\hat{p}_{\max(i,j)}}{\sum_{l=1}^n I(T_l > \min(t_{(i)}, t_{(j)})) \hat{p}_l}.$$

When $i > L_t + m_t$ When $i = L_t + 1, \dots, L_t + m_t$

$$w_i = \frac{1}{2m_t} \sum_{j=L_t+1}^{L_t+m_t} \frac{\hat{p}_i}{\sum_{l=1}^n I(T_l > t_j) \hat{p}_l} = \frac{1}{2m_t} \sum_{j=L_t+1}^{L_t+m_t} \frac{\hat{p}_i}{\sum_{l=j+1}^n \hat{p}_l} \leq \frac{1}{2m_t}.$$

$$\therefore 4m_t^2 * \max |w_i|^2 \leq 4$$

$$\begin{aligned}
w_i &= \frac{1}{2m_t} \left(\sum_{j < i} \frac{\hat{p}_i}{\sum_{l=1}^n I(T_l > t_j) \hat{p}_l} + \sum_{j > i} \frac{\hat{p}_j}{\sum_{l=1}^n I(T_l > t_i) \hat{p}_l} \right) = \\
&= \frac{1}{2m_t} \left(\sum_{j=L_t+1}^{i-1} \frac{\frac{\hat{p}_i}{n}}{\sum_{l=j+1}^n \hat{p}_l} + \sum_{j=i+1}^n \frac{\frac{\hat{p}_j}{n}}{\sum_{l=i+1}^n \hat{p}_l} \right) = \\
&= \frac{1}{2m_t} \left(\sum_{j=L_t+1}^{i-1} \frac{\frac{\hat{p}_i}{n}}{\sum_{l=j+1}^n \hat{p}_l} + 1 \right) \leq \frac{1}{2m_t} \times 2 = \frac{1}{m_t}.
\end{aligned}$$

Thus condition (i) is verified.

$$\begin{aligned}
4m_t^2 \sum_i w_i^2 &= 1^2 + \left(\frac{\frac{\hat{p}_{L_t+2}}{n}}{\sum_{l=L_t+2}^n \hat{p}_l} + 1 \right)^2 + \left(\frac{\frac{\hat{p}_{L_t+3}}{n}}{\sum_{l=L_t+2}^n \hat{p}_l} + \frac{\frac{\hat{p}_{L_t+3}}{n}}{\sum_{l=L_t+3}^n \hat{p}_l} + 1 \right)^2 + \\
&+ \dots + \left(\frac{\frac{\hat{p}_{L_t+m_t}}{n}}{\sum_{l=L_t+2}^n \hat{p}_l} + \frac{\frac{\hat{p}_{L_t+m_t}}{n}}{\sum_{l=L_t+3}^n \hat{p}_l} + \dots + \frac{\frac{\hat{p}_{L_t+m_t}}{n}}{\sum_{l=L_t+m_t}^n \hat{p}_l} + 1 \right)^2 + \\
&+ \left(\sum_{j=L_t+1}^{L_t+m_t} \frac{1}{\sum_{l=j+1}^n \hat{p}_l} \right)^2 \left(\sum_{i=L_t+m_t+1}^n \hat{p}_i^2 \right) \geq m_t
\end{aligned}$$

$$\begin{aligned}
\sum_{i \neq j} w_{ij}^2 &= \sum_{i=L_t+1}^n \sum_{\substack{j=L_t+1 \\ j \neq i}}^n w_{ij}^2 \quad \sum_i w_i^2 = \sum_{i=L_t+1}^n \left(\sum_{\substack{j=L_t+1 \\ j \neq i}}^n w_{ij} \right)^2 \\
4m_t^2 \sum_{i=L_t+1}^n \sum_{\substack{j=L_t+1 \\ j \neq i}}^n w_{ij}^2 &= \left(\frac{\hat{p}_{L_t+2}}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 + \left(\frac{\hat{p}_{L_t+3}}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 + \cdots + \left(\frac{\hat{p}_n}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 \\
&+ \left(\frac{\hat{p}_{L_t+2}}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 + \left(\frac{\hat{p}_{L_t+3}}{\sum_{l=L_t+3}^n \hat{p}_l} \right)^2 + \cdots + \left(\frac{\hat{p}_n}{\sum_{l=L_t+3}^n \hat{p}_l} \right)^2 + \\
&+ \left(\frac{\hat{p}_{L_t+3}}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 + \left(\frac{\hat{p}_{L_t+3}}{\sum_{l=L_t+3}^n \hat{p}_l} \right)^2 + \left(\frac{\hat{p}_{L_t+4}}{\sum_{l=L_t+3}^n \hat{p}_l} \right)^2 + \cdots + \left(\frac{\hat{p}_n}{\sum_{l=L_t+4}^n \hat{p}_l} \right)^2 + \\
&+ \cdots + \left(\frac{\hat{p}_{L_t+m_t}}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 + \cdots + \left(\frac{\hat{p}_{L_t+m_t}}{\sum_{l=L_t+m_t}^n \hat{p}_l} \right)^2 + \\
&+ \left(\frac{\hat{p}_{L_t+m_t+1}}{\sum_{l=L_t+m_t+1}^n \hat{p}_l} \right)^2 + \cdots + \left(\frac{\hat{p}_n}{\sum_{l=L_t+m_t+1}^n \hat{p}_l} \right)^2 + \\
&+ \left(\frac{\hat{p}_{L_t+m_t+1}}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 + \left(\frac{\hat{p}_{L_t+m_t+1}}{\sum_{l=L_t+3}^n \hat{p}_l} \right)^2 + \cdots + \left(\frac{\hat{p}_{L_t+m_t+1}}{\sum_{l=L_t+m_t+1}^n \hat{p}_l} \right)^2 + \\
&+ \cdots + \left(\frac{\hat{p}_n}{\sum_{l=L_t+2}^n \hat{p}_l} \right)^2 + \cdots + \left(\frac{\hat{p}_n}{\sum_{l=L_t+m_t+1}^n \hat{p}_l} \right)^2 = \\
&= \frac{2}{\left(\sum_{l=L_t+2}^n \hat{p}_l \right)^2} \left(\hat{p}_{L_t+2}^2 + \hat{p}_{L_t+3}^2 + \cdots + \hat{p}_n^2 \right) + \\
&+ \frac{2}{\left(\sum_{l=L_t+3}^n \hat{p}_l \right)^2} \left(\hat{p}_{L_t+3}^2 + \hat{p}_{L_t+4}^2 + \cdots + \hat{p}_n^2 \right) + \\
&+ \cdots + \frac{2}{\left(\sum_{l=L_t+m_t+1}^n \hat{p}_l \right)^2} \left(\hat{p}_{L_t+m_t+1}^2 + \hat{p}_{L_t+m_t+2}^2 + \cdots + \hat{p}_n^2 \right)
\end{aligned}$$

We define $u_k = \frac{2}{\left(\sum_{l=L_t+k}^n \hat{p}_l \right)^2} \left(\hat{p}_{L_t+k}^2 + \hat{p}_{L_t+k+1}^2 + \cdots + \hat{p}_n^2 \right)$, $k \in [2, m_t + 1]$

According to d'Alembert's ratio test: if $\sum u_n$ is a positive series and

$$\lim_{n \rightarrow \infty} \frac{u_{n+1}}{u_n} < 1$$

Then the series $\sum u_n$ converges, which indicates that $\sum u_n < \infty$ as $n \rightarrow \infty$

In our situation,

$$\begin{aligned} \frac{u_{k+1}}{u_k} &= \frac{\frac{1}{\left(\sum_{l=L_t+k+1}^n \hat{p}_l\right)^2} \left(\hat{p}_{L_t+k+1}^2 + \dots + \hat{p}_n^2\right)}{\frac{1}{\left(\sum_{l=L_t+k}^n \hat{p}_l\right)^2} \left(\hat{p}_{L_t+k}^2 + \dots + \hat{p}_n^2\right)} = \\ &= \left(\frac{\hat{p}_{L_t+k} + \hat{p}_{L_t+k+1} + \dots + \hat{p}_n}{\hat{p}_{L_t+k+1} + \dots + \hat{p}_n}\right)^2 \frac{\hat{p}_{L_t+k+1}^2 + \dots + \hat{p}_n^2}{\hat{p}_{L_t+k+1}^2 + \dots + \hat{p}_n^2} = \\ &= \left(1 + \frac{\hat{p}_{L_t+k}}{\sum_{l=L_t+k+1}^n \hat{p}_l}\right)^2 \left(1 - \frac{\hat{p}_{L_t+k}^2}{\sum_{l=L_t+k}^n \hat{p}_l^2}\right) = \\ &= 1 + \frac{2\hat{p}_{L_t+k}}{\sum_{l=L_t+k}^n \hat{p}_l} + \frac{\hat{p}_{L_t+k}^2}{\left(\sum_{l=L_t+k+1}^n \hat{p}_l\right)^2} - \frac{\hat{p}_{L_t+k}^2}{\sum_{l=L_t+k}^n \hat{p}_l^2} - \\ &\quad - \frac{\hat{p}_{L_t+k}^3}{\sum_{l=L_t+k+1}^n \hat{p}_l \sum_{l=L_t+k}^n \hat{p}_l^2} - \frac{\hat{p}_{L_t+k}^4}{\sum_{l=L_t+k+1}^n \hat{p}_l^2 \left(\sum_{l=L_t+k}^n \hat{p}_l\right)^2} \\ &< 1 \end{aligned}$$

Thus, the condition of d'Alembert's ratio test holds,

$$\begin{aligned} 4m_t^2 \sum_{i=L_t+1}^n \sum_{\substack{j=L_t+1 \\ j \neq i}}^n w_{ij}^2 &< \infty \quad \text{as } n \rightarrow \infty \\ \because \sum_i w_i^2 &\geq m_t \quad \therefore \frac{\sum_{i \neq j} w_{ij}^2}{\sum_i w_i^2} \rightarrow 0 \quad \text{when } m_t \rightarrow \infty \quad \text{as } n \rightarrow \infty \end{aligned}$$

Thus condition (ii) is verified.

The condition (iii) holds for $\psi_1(X_i, X_j) = \text{sgn}(j - i) \text{sgn}(R_i - R_j)$ since $|\psi_1(X_i, X_j)| \leq 1$. Finally,

$$V^{-1/2} \left[A^{\hat{MDS}}(t) - AUC(t) \right] \rightarrow_d N(0, 1)$$

References

- Bamber D. The area above the ordinal dominance graph and the area below the operating graph. *Journal of Mathematical Psychology*. 1975; 12:387–415.

- Breslow NE, Day NE. Statistical methods in cancer research. Vol. I – The analysis of case-control studies. IARC Scientific Publications. 1980; 32(32):5–338. [PubMed: 7216345]
- DeLong ER, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988; 44:837–845. [PubMed: 3203132]
- Ding J, Zhou H, Liu Y, Cai J, Longnecker MP. Estimating effect of environmental contaminants on women's subfecundity for the MoBa study data with an outcome-dependent sampling scheme. *Biostatistics*. 2014; 15:636–650. [PubMed: 24812419]
- Edelman M, Wang XF, Hodgson L, Cheney R, Baggstrom M, Sachdev T, Gajra A, Bertino E, Reckamp P, Ritter J, Vokes E. Baseline urinary PGE-M is a prognostic and predictive marker for Cox-2 inhibition in addition to standard chemotherapy for advanced non-small cell lung cancer (NSCLC): CALGB 30801 (alliance). *International Journal of Radiation Oncology, Biology, Physics*. 2014; 90:S39.
- Edelman M, Watson D, Wang XF, Morrison C, Kratzke R, Jewell S, Hodgson L, Mauer AM, Graziano SL, Masters GA, Bedor M, Green MJ, Vokes EE. Eicosanoid modulation in advanced lung cancer: COX-2 expression is a positive predictive factor for celecoxib + chemotherapy. *Journal of Clinical Oncology*. 2007; 26:848–855. [PubMed: 18281656]
- Etzioni R, Pepe M, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making*. 1999; 19:242–251. [PubMed: 10424831]
- Hanley JA, McNeil BJ. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
- Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*. 2000; 56:337–344. [PubMed: 10877287]
- Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. *Biometrics*. 2005; 61:92–105. [PubMed: 15737082]
- Morara M, Ryan L, Houseman A, Strauss W. Optimal design for epidemiological studies subject to designed missingness. *Lifetime Data Analysis*. 2007; 13:583–605. [PubMed: 18080755]
- Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*. 1988; 75:237–249.
- Owen AB. Empirical likelihood for confidence regions. *Annals of Statistics*. 1990; 18:90–120.
- Pepe, MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press; New York, NY: 2003.
- Qin J, Lawless JF. Empirical likelihood and general estimating equations. *Annals of Statistics*. 1994; 22:300–325.
- Saha-Chaudhuri P, Heagerty PJ. Nonparametric estimation of a time-dependent predictive accuracy curve. *Biostatistics*. 2013; 14:42–59. [PubMed: 22734044]
- Schildcrout JS, Rathouz PJ, Zelnick LR, Garbett SP, Heagerty PJ. Biased sampling designs to improve research efficiency: factors influencing pulmonary function over time in children with asthma. *The Annals of Applied Statistics*. 2015; 9:731–753. [PubMed: 26322147]
- Selen A, Dickinson PA, Mllertz A, Crison JR, Mistry HB, Cruaes MT, Martinez MN, Lennerns H, Wigal TL, Swinney DC, Polli JE. The biopharmaceutics risk assessment roadmap for optimizing clinical drug product performance. *Journal of Pharmaceutical Sciences*. 2014; 103:3377–3397. [PubMed: 25256402]
- Slate EH, Turnbull BW. Statistical models for longitudinal biomarkers of disease onset. *Statistics in Medicine*. 2000; 19:617–637. [PubMed: 10694740]
- Staniswalis JG. Local bandwidth selection for Kernel estimates. *Journal of the American Statistical Association*. 1989; 84:284–288.
- Strauss WJ, Ryan L, Morara M, Iroz-Elardo N, Davis M, Cupp M, Nishioka MG, Quackenboss J, Galke W, Zkaynak H, Scheidt P. Improving cost-effectiveness of epidemiological studies via designed missingness strategies. *Statistics in Medicine*. 2010; 29:1377–1387. [PubMed: 20527011]
- Van der Vaart, AW. *Asymptotic Statistics*. Cambridge University Press; Cambridge, UK: 2000.

- Wang XF, Ma JL, George SL. ROC curve estimation under test-result-dependent sampling. *Biostatistics*. 2013; 14:160–172. [PubMed: 22723502]
- Wang XF, Ma JL, George SL, Zhou HB. Estimation of AUC or partial AUC under test-result-dependent sampling. *Statistics in Biopharmaceutical Research*. 2012; 4:313–323. [PubMed: 23393612]
- Wang XF, Wu YG, Zhou HB. Outcome- and auxiliary-dependent subsampling and its statistical inference. *Journal of Biopharmaceutical Statistics*. 2009; 19:1132–1150. [PubMed: 20183468]
- Wang XF, Zhou HB. Design and inference for cancer biomarker study with an outcome and auxiliary-dependent subsampling. *Biometrics*. 2009; 66:502–511. [PubMed: 19508239]
- Wang XF, Zhou HB. A semiparametric empirical likelihood method for biased sampling schemes in epidemiologic studies with auxiliary covariates. *Biometrics*. 2006; 62:1149–1160. [PubMed: 17156290]
- Weaver MA, Zhou HB. An estimated likelihood method for continuous regression models with outcome-dependent sampling. *Journal of American Statistical Association*. 2005; 100:459–469.
- Williams PL, Seage GR, Van Dyke RB, Siberry GK, Griner R, Tassiopoulos K, Yildirim C, Read JS, Huo Y, Hazra R, Jacobson DL. A trigger-based design for evaluating the safety of in utero antiretroviral exposure in uninfected children of human immunodeficiency virus-infected mothers. *American Journal of Epidemiology*. 2012; 175:950–961. [PubMed: 22491086]
- Zhou HB, Weaver MA, Qin J, Longnecker MP, Wang MC. A semiparametric empirical likelihood method for data from an outcome dependent sampling scheme with a continuous outcome. *Biometrics*. 2002; 58:413–421. [PubMed: 12071415]
- Zhou XH. A nonparametric maximum likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias. *Biometrics*. 1996; 52:299–305. [PubMed: 8934599]

Figure 1.
 $\hat{A}^{MDS}$, $\hat{A}^{SRS}$ and their corresponding point-wise confidence intervals for the COX-2 data example.

Table 1

Bias simulation results under MDS.

$\log(t)$	-2	-1.5	-1	-0.5	0	0.5	1
$AUC(t)$	0.884	0.833	0.782	0.733	0.693	0.66	0.634
$\hat{A}^{SRS}$	0.86	0.823	0.79	0.761	0.732	0.711	0.693
Bias SRS	2.658%	1.204%	1.132%	3.783%	5.621%	7.766%	9.184%
$\hat{A}^{MDS}$	0.883	0.833	0.782	0.733	0.693	0.66	0.633
Bias MDS	0.118%	0.003%	0.016%	0.02%	0.045%	0.097%	0.064%

Notes: The simulations were conducted with bandwidth $= b \times n^{-0.2}$, $b^{MDS} = (2.8, 3.4, 3.7, 3.7, 4.7, 5, 5.3)$, $(\log(T), M) \sim \mathcal{N}(0, 0, 1, 1, -0.7)$, $(a_1, a_2) = [\mu M - \sigma M, \mu M + \sigma M]$, $n_S = (180, 60, 60, 60)$, and 20% censoring.

Table 2

Estimation of standard error under MDS with 2000 bootstrap simulations.

$\log(t)$	-2	-1.5	-1	-0.5	0	0.5	1
$AUC(t)$	0.884	0.833	0.782	0.733	0.693	0.66	0.634
$\hat{A}^{MDS}$	0.883	0.833	0.782	0.733	0.693	0.66	0.633
Bias	0.118%	0.003%	0.016%	0.02%	0.045%	0.097%	0.064%
SE	0.020	0.019	0.018	0.018	0.018	0.019	0.023
Bootstrap bias	0.073%	0.052%	0.072%	0.037%	0.068%	0.112%	0.085%
Bootstrap SE	0.017	0.017	0.016	0.017	0.017	0.019	0.021
95% CP	90.8%	94.0%	94.2%	94.6%	94.8%	94.3%	94.1%

Notes: The simulations were conducted with bandwidth = $b \times n^{-0.2}$, $b^{MDS} = (2.8, 3.4, 3.7, 3.7, 4.7, 5, 5.3)$, $(\log(T), M) \sim \mathcal{N}(0, 0, 1, 1, -0.7)$, $(a1, a2) = [\mu M - \sigma M, \mu M + \sigma M]$, $n_S = (180, 60, 60, 60)$ and 20% censoring.

Table 3

Efficiency comparison of MDS design over SRS design.

$\log(t)$	-2	-1.5	-1	-0.5	0	0.5	1
$AUC(t)$	0.884	0.833	0.782	0.733	0.693	0.66	0.634
$\hat{A}^{SRS}$ under SRS	0.878	0.828	0.78	0.731	0.692	0.662	0.633
Bias of $\hat{A}^{SRS}$ under SRS	0.687%	0.66%	0.159%	0.417%	0.062%	0.271%	0.135%
RMSE of $\hat{A}^{SRS}$ under SRS	0.04086	0.03519	0.03	0.02817	0.03315	0.03813	0.04076
$\hat{A}^{MDS}$ under MDS	0.883	0.833	0.782	0.733	0.693	0.66	0.633
Bias of $\hat{A}^{MDS}$ under MDS	0.118%	0.003%	0.016%	0.02%	0.045%	0.097%	0.064%
RMSE of $\hat{A}^{MDS}$ under MDS	0.02187	0.01909	0.01956	0.01837	0.02008	0.02277	0.02425

Notes: The simulations were conducted with bandwidth = $b \times n^{-0.2}$, $b^{SRS} = (1, 1, 1, 1, 1, 1, 1)$, $b^{MDS} = (2.8, 3.4, 3.7, 3.7, 4.7, 5, 5.3)$, $(\log(T), M) \sim N(0, 0, 1, 1, -0.7)$, $(a_1, a_2) = [\mu_M - \sigma_M, \mu_M + \sigma_M]$, $n_S = (180, 60, 60, 60)$, and 20% censoring.

Table 4

Performance of proposed estimator under different simulation scenarios.

		$\hat{A}^{MDS}$		$\hat{A}^{SRS}$		$\hat{A}^{SRS}$ under SRS	
$\log(t)$	AUC(t)	Bias	RMSE	Bias	RMSE	Bias	RMSE
$n_s = (180, 60, 60, 60)$							
−0.5	0.734	0.020%	0.018	5.127%	0.047	0.417%	0.028
0	0.693	0.045%	0.020	5.530%	0.049	0.062%	0.033
0.5	0.660	0.097%	0.023	3.474%	0.041	0.271%	0.038
1	0.634	0.064%	0.024	1.647%	0.044	0.135%	0.041
$n_s = (120, 80, 80, 80)$							
−0.5	0.734	0.242%	0.021	7.087%	0.058	0.066%	0.029
0	0.693	0.006%	0.021	6.714%	0.056	0.313%	0.031
0.5	0.660	0.176%	0.024	4.308%	0.045	0.532%	0.036
1	0.634	0.375%	0.026	1.583%	0.045	0.676%	0.046
$n_s = (240, 40, 40, 40)$							
−0.5	0.734	0.096%	0.019	3.595%	0.038	0.160%	0.029
0	0.693	0.237%	0.018	3.767%	0.040	0.034%	0.030
0.5	0.660	0.677%	0.020	2.828%	0.039	0.426%	0.036
1	0.634	0.159%	0.023	2.073%	0.048	0.225%	0.045
$n_s = (180, 30, 60, 90)$							
−0.5	0.734	1.076%	0.015	3.041%	0.027	0.036%	0.030
0	0.693	1.001%	0.028	3.372%	0.038	0.291%	0.030
0.5	0.660	0.987%	0.031	2.630%	0.042	0.423%	0.035
1	0.634	0.690%	0.032	2.224%	0.053	0.075%	0.045
40% censoring							
$n_s = (180, 60, 60, 60)$							
−0.5	0.734	0.573%	0.022	5.391%	0.050	0.041%	0.034
0	0.693	0.437%	0.022	5.423%	0.051	0.114%	0.035
0.5	0.660	0.100%	0.025	3.758%	0.049	0.596%	0.039
1	0.634	0.253%	0.028	1.396%	0.053	0.264%	0.055

Table 5

Median PFSs (months) and 95% CIs for CALGB 30203 and CALGB 30801.

		COX-2 score			
		[0,2)	[2,4)	[4,10]	All
C30203	Celecoxib	4.1 (2.8, 5.9)	3.2 (1.3, 7.0)	6.5 (4.4, 8.4)	5.5 (3.9, 6.4)
	Placebo	3.8 (2.3, 5.2)	8.8 (6.7, 10.8)	3.4 (0.7, 6.4)	4.1 (3.4, 6.1)
C30801	Celecoxib		4.8 (3.3, 6.7)	5.2 (4.1, 6.7)	5.2 (4.4, 6.0)
	Placebo		4.6 (3.8, 6.9)	5.5 (4.1, 6.9)	5.4 (4.3, 6.2)