Published in final edited form as:

Lifetime Data Anal. 2016 April ; 22(2): 280–298. doi:10.1007/s10985-015-9331-2.

A Log Rank Type Test in Observational Survival Studies with Stratified Sampling

Xiaofei Bai and Anastasios A. Tsiatis

Xiaofei Bai: xbai3@ncsu.edu; Anastasios A. Tsiatis: tsiatis@ncsu.edu

Abstract

In randomized clinical trials, the log rank test is often used to test the null hypothesis of the equality of treatment-specific survival distributions. In observational studies, however, the ordinary log rank test is no longer guaranteed to be valid. In such studies we must be cautious about potential confounders; that is, the covariates that affect both the treatment assignment and the survival distribution. In this paper, two cases were considered: the first is when it is believed that all the potential confounders are captured in the primary database, and the second case where a substudy is conducted to capture additional confounding covariates. We generalize the augmented inverse probability weighted complete case (AIPWCC) estimators for treatment-specific survival distribution proposed in Bai et al. (2013) and develop the log rank type test in both cases. The consistency and double robustness of the proposed test statistics are shown in simulation studies. These statistics are then applied to the data from the observational study that motivated this research.

Keywords

Cox proportional hazards model; Log rank test; Observational study; Stratified sampling; Survival analysis

1 Introduction

The ASCERT study, funded by the National Heart Lung and Blood Institute, was an observational study of patients with either two-vessel or three-vessel coronary artery disease. The patients in the study were treated either by surgical revascularization (coronary artery bypass grafting; CABG) or catheter-based revascularization (percutaneous coronary intervention; PCI). The goal of the ASCERT study is to compare these two treatment options for patients with coronary artery disease (Weintraub et al., 2012). Not all patients were followed until the endpoint of interest (all-cause mortality); accordingly, survival outcomes for these patients were censored. Moreover, a substudy was conducted to collect additional covariate information using a stratified random sample of patients from the main ASCERT study. Bai et al. (2013) used semiparametric methods to provide doubly-robust estimators of treatment-specific survival distributions using the data from the main study as well as taking advantage of the additional potential confounders from the substudy. Besides estimating treatment-specific survival distributions, the comparison of treatment-specific survival distributions is also of importance. In a randomized controlled clinical trial where the covariate distribution is balanced among different treatment groups, the log rank test is

most commonly used to test the null hypothesis of no difference between the treatment-specific survival distributions. In an observational study, however, the traditional log rank test is no longer valid due to possible confounding.

Xie and Liu (2005) proposed an inverse probability weighted log rank test to adjust for such confounding. They computed a weighted version of log rank statistics by substituting inverse probability weighted number of subject at risk and number of subject that die in each group. This estimator would provide valid inference if the propensity model is consistently estimated. Zhang and Schaubel (2012) compared treatment groups in term of the difference in restricted mean lifetime, which they referred as the average causal effect. This is a semiparametric estimator with a doubly-robust property. However, its consistency relies on the underlying assumption that the censoring time and survival time are independent conditional on the treatment. If the covariates also have effect on censoring which may likely be the case in observational studies, this method might be subject to bias.

In this paper, we propose a log rank type test statistic to compare treatment groups. With a nonparametric bootstrap estimator of the denominator, the resulting test statistic will be doubly-robust. Moreover, we generalize the test statistic into the case where a substudy is to be conducted to collect additional covariates to account for possible residual confounding.

In the next section, we derive the log rank test statistic using the data from the main study which is appropriate when the belief is that the collected covariates are enough to adjust for all potential confounding. In Section 3, we develop test statistics to deal with the case when we further conduct a stratified sampling substudy to collect additional potential confounding covariates. The performance of the proposed test statistics are demonstrated by simulation studies in Section 4. We then apply our method to the motivating example of ASCERT study to test for the treatment differences in CABG and PCI in Section 5. The paper ends with discussion and conclusion in Section 6.

2 Test Statistic for Main Study when all Potential Confounders are Captured

As in the motivating example of the ASCERT study, we are interested in the comparison of survival distributions in an observational study. In this article, we use the same notation as in Bai et al. (2013). We denote the two treatments used in the study as treatment 1 and treatment 0. For each individual i in the main study, we define the potential outcomes $(X_i, T_i^{(0)}, T_i^{(1)}, C_i^{(0)}, C_i^{(1)})$, $i = 1, \dots, N$, where X_i denotes a vector of baseline covariates, $T_i^{(j)}$ denotes the potential survival time and $C_i^{(j)}$ denotes the potential censoring time if individual i were given treatment (possibly contrary to fact) j , for $j = 0, 1$. In the case of the ASCERT study, all patients were followed from the date they entered the study until the time of data analysis and their survival status was fully ascertained during this period. Consequently, the potential censoring time was the time from entry into study until time of analysis which would be the same under both treatments. Therefore we have $C_i^{(1)} = C_i^{(0)} = C_i$ for $i = 1, \dots, N$. The treatment-specific survival distributions are defined as $S^{(j)}(u) = P(T^{(j)} > u)$ for $j = 0, 1$ and the main focus of this paper is to compare the overall survival distributions $S^{(1)}(u)$ and $S^{(0)}(u)$; specifically, to develop a test of the null hypothesis $H_0: S^{(0)}(u) = S^{(1)}(u)$, $u \geq 0$.

We also denote by Z_i the binary treatment assignment for patient i , where $Z_i = 1, 0$ and make the strong ignorability assumption (Rubin (1978)), or no unmeasured confounders assumption, that Z is conditionally independent of $T^{(j)}$ given X , denoted by $Z \perp\!\!\!\perp T^{(j)}|X$. We also make the usual assumption of non-informative censoring; namely, that $C \perp\!\!\!\perp T^{(j)}|X, Z$. Together these assumptions imply that

$$(Z, C) \perp\!\!\!\perp T^{(j)}|X. \quad (1)$$

We use this assumption in the remainder of this paper and refer to it as assumption (1). Besides assumption (1), we adopt the positivity assumption that $0 < P(Z = 1|X) < 1$ and the stable unit treatment value assumption (SUTVA) that one unit's outcomes are unaffected by another unit's treatment assignment.

In contrast to the potential outcomes, some of which may not be observed, the data that are observable can be summarized as $V_i = (Z_i, X_i, U_i, \delta_i), i = 1, \dots, N$, where, in addition to (Z_i, X_i) , already defined, we also observe for individual i their time to death or censoring $U_i = \min(T_i, C_i)$ and the failure indicator $\delta_i = I(T_i \leq C_i)$ where $T_i = Z_i T_i^{(1)} + (1 - Z_i) T_i^{(0)}$ (the consistency assumption); that is, the time to death on the assigned treatment.

In this section, we develop the log rank type statistic when the covariates collected in the main study are believed to capture all potential confounding. In Bai et al. (2013), we developed a locally-efficient doubly-robust estimator for the treatment-specific survival distributions $S^{(j)}(u) = P(T^{(j)} > u), u \geq 0, j = 0, 1$ as

$$\hat{S}^{(j)}(u) = N^{-1} \sum_{i=1}^N \phi_j^*(u; V_i) = N^{-1} \sum_{i=1}^N \left[\frac{(Z_i)^j (1 - Z_i)^{1-j} I(U_i \geq u)}{\{\pi(X_i)\}^j \{1 - \pi(X_i)\}^{1-j} K_c^{(j)}(u, X_i)} - \frac{(2j-1)\{Z_i - \pi(X_i)\}}{\{\pi(X_i)\}^j \{1 - \pi(X_i)\}^{1-j}} H^{(j)}(u, X_i) + \int_0^u \frac{(Z_i)^j (1 - Z_i)^{1-j}}{\{\pi(X_i)\}^j \{1 - \pi(X_i)\}^{1-j}} \frac{dM_c^{(j)}(r, X_i)}{K_c^{(j)}(r, X_i)} \frac{H^{(j)}(u, X_i)}{H^{(j)}(r, X_i)} dr \right], \quad (2)$$

where $\pi(X) = P(Z = 1|X)$ is the propensity score, $K_c^{(j)}(r, X) = P(C \geq r|X, Z=j)$ is the conditional survival function of the treatment specific censoring time given X , $dM_c^{(j)}(r, X)$ is the martingale increment for the censoring distribution, namely,

$$dM_c^{(j)}(r, X) = dN_c(r) - \lambda_c^{(j)}(r, X) Y(r) dr, \quad N_c(r) = I(U \leq r, \delta = 0), \quad Y(r) = I(U > r) \text{ and}$$

$\lambda_c^{(j)}(r, X) = \frac{-d \log K_c^{(j)}(r, X)}{dr}$ is the conditional hazard function for C given $Z = j$ and X , and $H^{(j)}(r, X) = P(T^{(j)} > r|X)$. Bai et al. (2013) also derived the estimator $\hat{S}^{(1)}(u) - \hat{S}^{(0)}(u)$ and its variance estimator allowing us to compare the survival distributions at a fixed time point u . We note that the monotonicity of the AIPWCC estimator $\hat{S}^{(j)}(u)$ for the survival distribution no longer holds. This is the price paid for increased efficiency. However, this does not affect the theoretical properties of the resulting log rank type test derived in this paper and with sufficiently large sample sizes our simulation studies have shown that the deviation from monotonicity of the AIPWCC estimator is slight.

In a randomized study with data $(Z_i, U_i, \delta_i), i = 1, \dots, N$, the log rank test takes the form of

$$\int W_N(u) \{d\hat{\Lambda}_1(u) - d\hat{\Lambda}_0(u)\},$$

where

$Y_1(u) = \sum_{i=1}^N Z_i Y_i(u)$, $Y_0(u) = \sum_{i=1}^N (1 - Z_i) Y_i(u)$, $N_1(u) = \sum_{i=1}^N Z_i N_i(u)$, $N_0(u) = \sum_{i=1}^N (1 - Z_i) N_i(u)$, $d\hat{\Lambda}_j(u) = dN_j(u)/Y_j(u)$ is the usual treatment specific Nelson-Aalen estimator for the hazard function of the survival distribution for treatment $j = 0, 1$, and $W_N(u) = N^{-1} \frac{Y_1(u)Y_0(u)}{Y_1(u) + Y_0(u)}$. We note that we normalized $W_N(u)$ so that $W_N(u)$ converges in probability to

$w(u) = \frac{P(U \geq u | Z=1)\xi \times P(U \geq u | Z=0)(1-\xi)}{P(U \geq u | Z=1)\xi + P(U \geq u | Z=0)(1-\xi)}$, where $\xi = P(Z = 1)$. By analogy, we propose to use $\int W_N(u) \{d\hat{\Lambda}^{(1)}(u) - d\hat{\Lambda}^{(0)}(u)\}$, where $d\hat{\Lambda}^{(j)}(u) = -d\hat{S}^{(j)}(u)/\hat{S}^{(j)}(u)$, $j = 0, 1$, to develop a log rank type test of the null hypothesis of treatment equality in observational studies. Here the estimator for the treatment-specific survival distributions $\hat{S}^{(j)}(u)$ uses the doubly robust estimator (2). Define

$$T_N = \int W_N(u) \{d\hat{\Lambda}^{(1)}(u) - d\hat{\Lambda}^{(0)}(u)\} = \int W_N(u) \left\{ \frac{d\hat{S}^{(0)}(u)}{\hat{S}^{(0)}(u)} - \frac{d\hat{S}^{(1)}(u)}{\hat{S}^{(1)}(u)} \right\},$$

and

$$T_N^* = \int W_N(u) \left[\left\{ \frac{d\hat{S}^{(0)}(u)}{\hat{S}^{(0)}(u)} - \frac{dS^{(0)}(u)}{S^{(0)}(u)} \right\} - \left\{ \frac{d\hat{S}^{(1)}(u)}{\hat{S}^{(1)}(u)} - \frac{dS^{(1)}(u)}{S^{(1)}(u)} \right\} \right],$$

under the null hypothesis of $H_0: S^{(1)}(u) = S^{(0)}(u)$, we have $T_N = T_N^*$, hence

$\text{Var}(T_N) = \text{Var}(T_N^*)$. Moreover,

$$\begin{aligned} T_N^* &= \int W_N(u) \left[\left\{ \frac{d\hat{S}^{(0)}(u)}{\hat{S}^{(0)}(u)} - \frac{dS^{(0)}(u)}{S^{(0)}(u)} + \frac{dS^{(0)}(u)}{\hat{S}^{(0)}(u)} - \frac{dS^{(0)}(u)}{S^{(0)}(u)} \right\} \right. \\ &\quad \left. - \left\{ \frac{d\hat{S}^{(1)}(u)}{\hat{S}^{(1)}(u)} - \frac{dS^{(1)}(u)}{S^{(1)}(u)} + \frac{dS^{(1)}(u)}{\hat{S}^{(1)}(u)} - \frac{dS^{(1)}(u)}{S^{(1)}(u)} \right\} \right] \\ &= \int \left[\frac{W_N(u)}{\hat{S}^{(0)}(u)} d \left\{ \hat{S}^{(0)}(u) - S^{(0)}(u) \right\} - \frac{W_N(u)dS^{(0)}(u)}{\hat{S}^{(0)}(u)S^{(0)}(u)} \left\{ \hat{S}^{(0)}(u) - S^{(0)}(u) \right\} \right] \\ &\quad - \left[\frac{W_N(u)}{\hat{S}^{(1)}(u)} d \left\{ \hat{S}^{(1)}(u) - S^{(1)}(u) \right\} - \frac{W_N(u)dS^{(1)}(u)}{\hat{S}^{(1)}(u)S^{(1)}(u)} \left\{ \hat{S}^{(1)}(u) - S^{(1)}(u) \right\} \right]. \end{aligned}$$

Since $\hat{S}^{(j)}(u) - S^{(j)}(u) = \frac{1}{N} \sum_{i=1}^N \phi_j(u; V_i)$, where $\phi_j(u; V_i) = \phi_j^*(u; V_i) - S^{(j)}(u)$, then T_N^* can be approximated by T_N^* given by

$$\hat{T}_N^* = \frac{1}{N} \sum_{i=1}^N \int \left[\frac{W_N(u)}{\hat{S}^{(0)}(u)} d\phi_0(u; V_i) - \frac{W_N(u)d\hat{S}^{(0)}(u)}{\{\hat{S}^{(0)}(u)\}^2} \phi_0(u; V_i) \right] - \left[\frac{W_N(u)}{\hat{S}^{(1)}(u)} d\phi_1(u; V_i) - \frac{W_N(u)d\hat{S}^{(1)}(u)}{\{\hat{S}^{(1)}(u)\}^2} \phi_1(u; V_i) \right],$$

in the sense that $N^{1/2}(T_N^* - \hat{T}_N^*)$ converges in probability to zero. For each subject i , we first calculate

$$\psi_i = \int \left[\frac{W_N(u)}{\hat{S}^{(0)}(u)} d\phi_0(u; V_i) - \frac{W_N(u) d\hat{S}^{(0)}(u)}{\{\hat{S}^{(0)}(u)\}^2} \phi_0(u; V_i) \right] - \left[\frac{W_N(u)}{\hat{S}^{(1)}(u)} d\phi_1(u; V_i) - \frac{W_N(u) d\hat{S}^{(1)}(u)}{\{\hat{S}^{(1)}(u)\}^2} \phi_1(u; V_i) \right], i=1, \dots, N.$$

Then we compute the empirical variance of ψ_i ,

$$\text{Var}(\psi_i) = \frac{1}{N-1} \sum_{i=1}^N \left(\psi_i - \frac{1}{N} \sum_{j=1}^N \psi_j \right)^2.$$

Finally, we use $\text{Var}(\psi_i)/N$ as the estimator to $\widehat{\text{Var}}(\hat{T}_N^*)$. The log rank type statistic takes the form

$$G = \frac{T_N}{\sqrt{\widehat{\text{Var}}(\hat{T}_N^*)}}, \quad (3)$$

which asymptotically follows a standard normal distribution $N(0, 1)$ under the null hypothesis of $H_0: S^{(1)}(u) = S^{(0)}(u)$. Hence, the null hypothesis is rejected at α level when $|G| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is $1 - \alpha/2$ quantile of standard normal distribution.

In Bai et al. (2013), it is shown that $\hat{S}^{(j)}(u)$, $j = 0, 1$, is doubly-robust; that is, if either the estimator for the coarsening defined through the estimator for $P(Z = 1|X) = \pi(X)$ and the estimator for $P(C \geq r|Z=j, X) = K_c^{(j)}(r, X)$ are both consistent or the estimator for $P(T^{(j)} r|X) = H^{(j)}(r, X)$ is consistent, then (2) will be a consistent estimator for $S^{(j)}(u)$. Hence, the numerator T_N is also a doubly-robust estimator of $\int w(u) \{dA_1(u) - dA_0(u)\}$; hence has expectation zero under the null hypothesis if either $\pi(X)$ and $K_c^{(j)}(r, X)$ are both consistently estimated or $H^{(j)}(r, X)$ is consistently estimated. The denominator estimator $\sqrt{\widehat{\text{Var}}(\hat{T}_N^*)}$, however, is not guaranteed to be valid under model misspecification. If the propensity score $\pi(X)$ and the conditional censoring distribution $K_c^{(j)}(r, X)$ were consistently estimated whereas, the conditional survival distribution $H^{(j)}(r, X)$ was not, then the variance estimator would be conservative (Hubbard et al. (2000); Tsiatis (2006) Theorem 9.1). The resulting

test statistic $\frac{T_N}{\sqrt{\widehat{\text{Var}}(\hat{T}_N^*)}}$ would follow a normal distribution with mean 0 and variance less than 1. On the other hand, when the estimator for $H^{(j)}(r, X)$ is consistent and either of the

estimators for $\pi(X)$ or $K_c^{(j)}(r, X)$ are not, then the resulting denominator $\sqrt{\widehat{\text{Var}}(\hat{T}_N^*)}$ may be biased and there is no theoretical result on the direction or extent of the bias. In such cases, it is suggested to use a bootstrap estimator of the asymptotic variance. To be more specific, we resample the dataset $(Z_i, X_i, U_i, \dots)$, $i = 1, \dots, n$, with replacement B times. In each bootstrap

replicate, we compute the numerator $T_N^{(b)}$, $b = 1, \dots, B$. We then compute the sample standard deviation of $T_N^{(b)}$; $b = 1, \dots, B$, and denote that by $\sqrt{\text{Var}\{T_N^{(b)}\}}$. We propose the

bootstrap test statistics as $G_{bootstrap} = \frac{T_N}{\sqrt{\text{Var}\{T_N^{(b)}\}}}$. As shown later in our simulation studies in Section 4, the bootstrap version of these test statistics have good asymptotic properties under varying model misspecification.

Similar to the ordinary log rank test, under local alternatives, these test statistics follow a normal distribution with a non-centrality parameter and variance 1. We compare the performance of this statistic with other competing methods in terms of power under various alternatives in the simulation studies in Section 4.

One thing worth pointing out is that both the numerator and denominator of (3) involve the reciprocal of $\hat{S}^{(j)}(u)$, $j = 0, 1$, which tends to be unstable at the tail of survival curve. Therefore, in order to lessen this problem one may truncate the range of integration of our test statistic. The validity of the distribution of this truncated test statistic will still hold under the null hypothesis. However, with truncation there is potential loss of efficiency by not using all the survival data. Thus, we need to evaluate the balance between increased stability versus loss of efficiency in the choice of the truncation point. This issue will also be examined in our simulation studies.

3 Test Statistics in Stratified Sampling with Additional Information from a Subsample

As discussed in Bai et al. (2013), the inference derived in Section 2 will be valid only when the critical no unmeasured confounder assumption $Z \perp\!\!\!\perp T^{(j)}|X$ holds for covariates collected in the main study. If this assumption is questionable then one may decide to collect additional covariate information in a substudy in order to make this assumption tenable. This indeed was the case for the ASCERT study where additional covariates were collected using a stratified sampling design. We now show how one can construct a valid log rank type test for comparing treatment-specific survival distributions where a substudy using a stratified sampling design to collect such additional covariates was used.

Using the same notation as in Bai et al. (2013), X_1 denotes the covariates collected in the main study and X_2 denotes the additional covariates collected on the subsample. Use $F_i = (U_i, Z_i, X_{1i}, X_{2i})$, $i = 1, \dots, N$ to denote the “full data” and

$$\begin{aligned} \phi_j^*(u; F_i) = & \left[\frac{(Z_i)^j (1-Z_i)^{1-j} I(U_i \geq u)}{\{\pi(X_i)\}^j \{1-\pi(X_i)\}^{1-j} K_c^{(j)}(u, X_i)} \right. \\ & - \frac{(2j-1) \{Z_i - \pi(X_i)\}}{\{\pi(X_i)\}^j \{1-\pi(X_i)\}^{1-j}} H^{(j)}(u, X_i) \\ & \left. + \int_0^u \frac{(Z_i)^j (1-Z_i)^{1-j}}{\{\pi(X_i)\}^j \{1-\pi(X_i)\}^{1-j}} \frac{dM_c^{(j)}(r, X_i)}{K_c^{(j)}(r, X_i)} \frac{H^{(j)}(u, X_i)}{H^{(j)}(r, X_i)} \right] \end{aligned}$$

to denote the i -th element of estimating equation that would have been used to estimate $S^{(j)}(u)$ if we had access to all the full data in the main study. This is similar to the estimating equation (2) but using both X_1 and X_2 . As in the ASCERT study, a subsample may be obtained using a stratified sampling design in which K strata are identified based on (Z, X_1) .

Let ε denote the stratum indicator taking values $1, \dots, K$ such that $N_k = \sum_{i=1}^N I(\varepsilon_i = k)$ is the number of subjects in stratum k , $k = 1, \dots, K$, then from stratum k , a fixed (by design) number of subjects n_k are sampled from the N_k subjects at random without replacement. As shown in Bai et al. (2013), the doubly-robust estimator for the treatment-specific survival probability at time point u is:

$$\hat{S}_{start}^{(j)}(u) = \frac{1}{N} \sum_{i=1}^N \left[\frac{R_i}{\eta(\varepsilon_i)} \hat{\phi}_j^*(u; F_i) - \left\{ \frac{R_i - \eta(\varepsilon_i)}{\eta(\varepsilon_i)} \right\} h^{(j)}(u; V_i) \right], \quad (4)$$

and its variance can be estimated by the summation of

$$N^{-1} \left[\sum_{k=1}^K p_k \mu_k^2(\phi_j) - \left\{ \sum_{k=1}^K p_k \mu_k(\phi_j) \right\}^2 \right] \quad (5)$$

and

$$\sum_{k=1}^K \left(\frac{N_k}{N} \right)^2 \left[\left\{ \text{Var}_k(\phi_j) - 2\text{Cov}_k(\phi_j, h^{(j)}) + \text{Var}_k(h^{(j)}) \right\} \left(\frac{1}{n_k} - \frac{1}{N_k} \right) + \text{Var}_k(\phi_j) \left(\frac{1}{N_k} \right) \right], \quad (6)$$

where R_i is the subsample indicator, $\eta(\varepsilon_i) = P(R_i = 1 | \varepsilon_i)$, $h^{(j)}(u; V_i) = E\{\phi_j^*(F_i) | V_i, \varepsilon_i\}$, $p_k = P(\varepsilon = k)$, $\phi_j = \phi_j^* - S^{(j)}(u)$, $\mu_k(\phi_j) = E[\phi_j \{F, S^{(j)}(u)\} | \varepsilon = k]$, $\mu_k(h^{(j)}) = E[h^{(j)}(V) | \varepsilon = k]$, $\text{Var}_k(\phi_j) = \text{Var}[\phi_j \{F, S^{(j)}(u)\} | \varepsilon = k]$, $\text{Var}_k(h^{(j)}) = \text{Var}\{h^{(j)}(V) | \varepsilon = k\}$, and $\text{Cov}_k(\phi_j, h^{(j)}) = \text{Cov}[\phi_j \{F, S^{(j)}(u)\}, h^{(j)}(V) | \varepsilon = k]$. Note that $h^{(j)}(u; V_i) = E\{\phi_j^*(F_i) | V_i, \varepsilon_i\}$, also referred to as the optimal h function or h^{opt} , is chosen such that the variance of $\hat{S}^{(j)}(u)$ is minimized. Such efficiency is achieved by recovering information from observations in the main study but not selected in the subsample. A simple and effective estimating strategy of h^{opt} is proposed in Bai et al. (2013), suggesting that we fit a simple linear regression of $\phi_j^*(F)$ on $\phi_j^*(V)$ by strata. We adopt this estimation method in the simulation and data analysis section later.

As in the case of the main study, the numerator of the test statistic is given by

$$T_{N, strat} = \int W_N(u) \left\{ \frac{d\hat{S}_{strat}^{(0)}(u)}{\hat{S}_{strat}^{(0)}(u)} - \frac{d\hat{S}_{strat}^{(1)}(u)}{\hat{S}_{strat}^{(1)}(u)} \right\}.$$

Under the null hypothesis, we have

$$= \int \left[\frac{W_N(u)}{\hat{S}_{strat}^{(0)}(u)} d \left\{ \hat{S}_{strat}^{(0)}(u) - S^{(0)}(u) \right\} - \frac{W_N(u) dS^{(0)}(u)}{\hat{S}_{strat}^{(0)}(u) S^{(0)}(u)} \left\{ \hat{S}_{strat}^{(0)}(u) - S^{(0)}(u) \right\} \right] - \left[\frac{W_N(u)}{\hat{S}_{strat}^{(1)}(u)} d \left\{ \hat{S}_{strat}^{(1)}(u) - S^{(1)}(u) \right\} - \frac{W_N(u) dS^{(1)}(u)}{\hat{S}_{strat}^{(1)}(u) S^{(1)}(u)} \left\{ \hat{S}_{strat}^{(1)}(u) - S^{(1)}(u) \right\} \right]$$

Hence,

$$\begin{aligned} \hat{T}_{N, strat}^* &= \frac{1}{N} \sum_{i=1}^N \left(\frac{R_i}{\eta(\varepsilon_i)} \int \left[\frac{W_N(u)}{\hat{S}_{strat}^{(0)}(u)} d\phi_0(u; F_i) - \frac{W_N(u) d\hat{S}_{strat}^{(0)}(u)}{\{\hat{S}_{strat}^{(0)}(u)\}^2} \phi_0(u; F_i) \right] \right. \\ &\quad \left. - \left[\frac{W_N(u)}{\hat{S}_{strat}^{(1)}(u)} d\phi_1(u; F_i) - \frac{W_N(u) d\hat{S}_{strat}^{(1)}(u)}{\{\hat{S}_{strat}^{(1)}(u)\}^2} \phi_1(u; F_i) \right] \right. \\ &\quad \left. - \frac{R_i - \eta(\varepsilon_i)}{\eta(\varepsilon_i)} \int \left[\frac{W_N(u)}{\hat{S}_{strat}^{(0)}(u)} dh^{(0)}(u; V_i) - \frac{W_N(u) d\hat{S}_{strat}^{(0)}(u)}{\{\hat{S}_{strat}^{(0)}(u)\}^2} h^{(0)}(u; V_i) \right] \right. \\ &\quad \left. - \left[\frac{W_N(u)}{\hat{S}_{strat}^{(1)}(u)} dh^{(1)}(u; V_i) - \frac{W_N(u) d\hat{S}_{strat}^{(1)}(u)}{\{\hat{S}_{strat}^{(1)}(u)\}^2} h^{(1)}(u; V_i) \right] \right). \end{aligned}$$

Notice that $\hat{T}_{N, strat}^*$ is of the same form as the AIPWCC stratified estimator (4), its variance can also be estimated by adding up the estimator of the variance of the conditional expectation terms like (5) and the estimator of the expectation of the conditional variance term like (6). To be specific, for each subject i , we first calculate

$$\psi_{i, strat} = \frac{R_i}{\eta(\varepsilon_i)} \zeta_i^* - \frac{R_i - \eta(\varepsilon_i)}{\eta(\varepsilon_i)} \gamma_i,$$

where

$$\zeta_i^* = \int \left[\frac{W_N(u)}{\hat{S}_{strat}^{(0)}(u)} d\phi_0(u; F_i) - \frac{W_N(u) d\hat{S}_{strat}^{(0)}(u)}{\{\hat{S}_{strat}^{(0)}(u)\}^2} \phi_0(u; F_i) \right] - \left[\frac{W_N(u)}{\hat{S}_{strat}^{(1)}(u)} d\phi_1(u; F_i) - \frac{W_N(u) d\hat{S}_{strat}^{(1)}(u)}{\{\hat{S}_{strat}^{(1)}(u)\}^2} \phi_1(u; F_i) \right],$$

and

$$\gamma_i = \int \left[\frac{W_N(u)}{\hat{S}_{strat}^{(0)}(u)} dh^{(0)}(u; V_i) - \frac{W_N(u) d\hat{S}_{strat}^{(0)}(u)}{\{\hat{S}_{strat}^{(0)}(u)\}^2} h^{(0)}(u; V_i) \right] - \left[\frac{W_N(u)}{\hat{S}_{strat}^{(1)}(u)} dh^{(1)}(u; V_i) - \frac{W_N(u) d\hat{S}_{strat}^{(1)}(u)}{\{\hat{S}_{strat}^{(1)}(u)\}^2} h^{(1)}(u; V_i) \right].$$

Then the variance $\widehat{\text{Var}}(\hat{T}_{N, strat}^*)$ can be estimated by the summation of

$$N^{-1} \left[\sum_{k=1}^K p_k \mu_k^2(\zeta) - \left\{ \sum_{k=1}^K p_k \mu_k(\zeta) \right\}^2 \right]$$

and

$$\sum_{k=1}^K \left(\frac{N_k}{N} \right)^2 \left[\{ \text{Var}_k(\zeta) - 2\text{Cov}_k(\zeta, \gamma) + \text{Var}_k(\gamma) \} \left(\frac{1}{n_k} - \frac{1}{N_k} \right) + \text{Var}_k(\zeta) \left(\frac{1}{N_k} \right) \right],$$

where $\zeta_i = \zeta_i^* - \hat{T}_{N, \text{strat}}^*$, $\mu_k(\zeta) = \sum_{\{i: \varepsilon_i = k, R_i = 1\}} \zeta_i / n_k$, $\mu_k(\gamma) = \sum_{\{i: \varepsilon_i = k\}} \gamma_i / N_k$, $\text{Var}_k(\zeta) = \text{Var}_{\{i: \varepsilon_i = k, R_i = 1\}}(\zeta_i)$, $\text{Var}_k(\gamma) = \text{Var}_{\{i: \varepsilon_i = k\}}(\gamma_i)$ and $\text{Cov}_k(\zeta, \gamma) = \text{Cov}_{\{i: \varepsilon_i = k, R_i = 1\}}(\zeta_i, \gamma_i)$. The final log rank type statistic in the stratified sampling framework is given by

$$G_{\text{strat}} = \frac{T_{N, \text{strat}}}{\sqrt{\widehat{\text{Var}}(\hat{T}_{N, \text{strat}}^*)}}. \quad (7)$$

The discussion at the end of Section 2 also applies to the case with stratified sampling: if all models are correctly specified, the test statistics follow a standard normal distribution under the null hypothesis and follow a noncentral normal distribution with variance 1 under the alternative hypothesis. Misspecification of the regression model $H^{(j)}(r, X)$ would result in a test statistic with variance less than 1; misspecification of the propensity model $\pi(X)$ and/or $K_c^{(j)}(r, X)$ would affect the variance of the test statistic in an unknown direction. Similar to the case of the main study, the nonparametric bootstrap procedure might provide a more accurate estimator of the asymptotic variance. One way to conduct the bootstrap sampling is to randomly sample with replacement the main study subjects and substudy subjects separately. That is, given a bootstrap number B , for each bootstrap replicate $b = 1, \dots, B$, a

total of $\sum_{i=1}^N R_i$ subjects are sampled with replacement among all subject with $R_i = 1$ and a total of $\sum_{i=1}^N (1 - R_i)$ subjects are sampled with replacement among all subject with $R_i = 0$. Combining the new N subjects sampled from the b -th bootstrap replicate, we compute

$T_{N, \text{strat}}^{(b)}$. The bootstrap test statistics in stratified sampling take would take the form

$$G_{\text{strat}, \text{bootstrap}} = \frac{T_{N, \text{strat}}}{\sqrt{\text{Var}\{T_{N, \text{strat}}^{(b)}\}}}, \text{ where } \sqrt{\text{Var}\{T_{N, \text{strat}}^{(b)}\}} \text{ is the sample standard deviation of}$$

$T_{N, \text{strat}}^{(b)}$, $b = 1, \dots, B$. As shown in the simulation study, the bootstrap version of test performs better compared to (7) in the case of misspecification. Also, appropriate truncation could result in more stable performance with possible slight efficiency loss.

4 Simulations

4.1 Simulation 1: Main Study Only

In this section, we conduct a simulation study on the performance of test statistics (3) under both the null hypothesis and various alternative hypotheses. Different type of misspecification scenarios, including misspecification of propensity models and of regression models, are also considered.

In this simulation, we generated 500 replications and within each replicate, $N = 2,000$ observations are generated as follows: $B \sim \text{Bernoulli}(0.5)$, $W \sim N(0, 1)$ and $X_2 \sim N(0, 1)$, mutually independent. We denote by $X_1 = (B, W, W^2)^T$, and $X = (X_1^T, X_2^T)^T$. The treatment assignment propensity is generated by $P(Z=1|X) = \frac{\exp(0.1B+0.1W+0.5W^2+0.5X_2)}{1+\exp(0.1B+0.1W+0.5W^2+0.5X_2)}$, and the treatment-specific hazard functions are given by $\lambda_{Z=1}(t|X) = \lambda_{Z=0}(t|X) = \exp(-0.5 + 0.1B + 0.1W + 0.5W^2 + 0.5X_2)$. We also generated an independent censoring variable C as $\text{Uniform}(0, 4)$. The censoring rate is approximately 28%. This is a scenario under the null hypothesis and we expect the test statistics (3) to follow a standard normal distribution.

With all models, $\pi(X)$, $K_c^{(j)}(r, X)$ and $H^{(j)}(r, X)$, being correctly specified, the performance of the log rank type test statistics (3) is summarized in Table 1. It is expected that the test statistics follow the standard normal distribution and the type I error rate is indeed close to 0.05, with or without truncation. We also computed the ordinary log rank test and because of confounding, this test does not provide valid inference and rejects the null hypothesis with probability 1. The mean of ordinary log rank test is 7.05 and the standard deviation is 0.99.

If we misspecify the propensity model by leaving out the W^2 term in $\pi(X)$ and keeping the other two models $K_c^{(j)}(r, X)$ and $H^{(j)}(r, X)$ correctly specified, the performance of log rank type test statistics (3) is summarized in Table 2. Theoretically, in this scenario, the test statistics should still follow the normal distribution with mean 0 but not necessarily a variance of one. In this scenario the variance estimator was close to one and hence the nominal level of significance was close to 0.05, however, this is not guaranteed, at least theoretically, to occur in general. Table 3 presents the bootstrap version of test statistics $G_{bootstrap}$ with $B = 100$ bootstrap resampling under wrong propensity model. We see after appropriate truncation, the bootstrap test provides nominal level of significance close to 0.05.

We also considered misspecification of the regression models by leaving out the W^2 term in $H^{(j)}(r, X)$ and keeping the propensity models $\pi(X)$ and $K_c^{(j)}(r, X)$ correctly specified, the performance of log rank type test statistics (3) is summarized in Table 4. Theoretically, in this scenario, the test statistics should still follow the normal distribution with mean 0 and variance less than 1, which is indeed supported by our simulation results. The bootstrap version of test statistics $G_{bootstrap}$ with $B = 100$ bootstrap resampling under wrong regression model is presented in Table 5. By using the bootstrap variance, the resulting log rank test statistic $G_{bootstrap}$ provides nominal significance level close to 0.05, even when the regression model is misspecified.

Under the alternative hypothesis of $\lambda_{Z=1}(t|X) = \exp(-\beta_0^{(1)} + 0.1B + 0.1W + 0.5W^2 + 0.5X_2)$, and $\lambda_{Z=0}(t|X) = \exp(-0.5 + 0.1B + 0.1W + 0.5W^2 + 0.5X_2)$, the power performance of the test statistics (3) is summarized in Table 6. The censoring rates are approximately 29% for $\beta_0^{(1)} = 0.6$ and 30% for $\beta_0^{(1)} = 0.7$. The model used to generate the data was a proportional hazards regression model with covariates X and Z , and here the null hypothesis would correspond to $\beta_0^{(1)} = 0.5$. Consequently, for this model the most powerful test of the null

hypothesis versus alternatives $\beta_0^{(1)} \neq 0.5$ is obtained by fitting a Cox proportional hazard model with covariate X and Z to the data and using the estimated coefficient of Z divided by its estimated standard deviation as the gold standard test statistic. The power of this gold standard test statistic is 0.376 when $\beta_0^{(1)}=0.6$ and 0.906 when $\beta_0^{(1)}=0.7$. This indicates that (3) is comparable to the optimal model based gold standard test.

4.2 Simulation 2: With Substudy

In this section, we consider the performance of test statistics (7) under both null and alternative hypotheses. Similar to the previous study, we also compute the test under different misspecification scenarios, including misspecification of the propensity models and regression models.

The data generating mechanism is the same as the previous studies, where $X_1 = (B, W, W^2)^T$ is collected on $N = 2000$ observations in the main study and X_2 is only observable for a random subsample of size $n_{11} = n_{12} = n_{01} = n_{02} = 300$, where n_{jk} denotes the number of subjects in stratum $Z = j$ and $B = k - 1$, for $j = 0, 1, k = 1, 2$. For this scenario we also computed the test statistics (3) using only X_1 and not adjusting for the residual confounding of X_2 . As expected, the resulting test statistics were biased leading to a rejection rate of roughly 0.95 under the null hypothesis for a test at the nominal 0.05 level.

In order to get consistent statistics, we need to adjust for additional con-founding captured by the substudy, X_2 . Besides (7), which will be referred to as AIPWCC stratified test statistics, we compute similar test statistics with no augmentation terms. That is, we compute (7) by substituting $h^{(j)}(u; V) = 0, j = 0, 1$ in both the numerator and denominator.

This estimator, referred to as weighted test statistics, $G_{strat}^{weighted}$, only include the subjects being selected into the substudy and we use information collected on them from both the main study and the substudy. Hence, it is expected to be less efficient than the AIPWCC stratified test statistics. Similar to $G_{strat,bootstrap}$, a bootstrap version of weighted test statistics, denoted by $G_{strat,bootstrap}^{weighted}$, could be computed by using the sample standard deviation as the denominator. The performance of $G_{strat,bootstrap}$ and $G_{strat,bootstrap}^{weighted}$ will be compared in the case of model misspecification.

When all models, $\pi(X)$, $K_c^{(j)}(r, X)$ and $H^{(j)}(r, X)$, are correctly specified, the performance of AIPWCC stratified test statistics and weighted test statistics are summarized in Table 7. Both test statistics follows the standard normal distribution and the type I error rate is close to the nominal 0.05 level, regardless of the degree of truncation. Similar to the case of main study only, the ordinary log rank test statistic does not account for the confounding appropriately and leads to biased result with mean 7.00 and standard deviation 1.01.

When the propensity model is misspecified by leaving out the W^2 term in $\pi(X)$ and the other two models $K_c^{(j)}(r, X)$ and $H^{(j)}(r, X)$ are correctly specified, the performance of the AIPWCC stratified test statistics and weighted test statistics are summarized in Table 8. Theoretically, in this scenario, the test statistics should still follow a normal distribution with mean 0 but with variance that may not equal one; however, our simulation results show that

the variance of both test statistics are close to 1 in this setting leading to nominal significance levels close to 0.05. Table 9 presents the result of the bootstrap test statistics $G_{strat,bootstrap}$ and $G_{strat,bootstrap}^{weighted}$. Both provide valid inference with nominal significance level.

If we misspecify the regression models by leaving out the W^2 term in $H^{(j)}(r, X)$ and keep the propensity models $\pi(X)$ and $K_c^{(j)}(r, X)$ correctly specified, the performance of the AIPWCC stratified test statistics and weighted test statistics are summarized in Table 10. In this scenario, both test statistics should still follow the normal distribution with mean 0 and variance less than 1. In particular, the variance of the AIPWCC stratified test statistics are larger than the weighted test statistics. This indeed is what we would expect. The bootstrap test statistics $G_{strat,bootstrap}$ and $G_{strat,bootstrap}^{weighted}$ are summarized in Table 11. Compared with Table 10, we see the bootstrap version of tests are more accurate. Hence, it would be suggested to perform the bootstrap test when there exists the possibility of misspecification.

Under the alternative hypothesis of $\lambda_{Z=1}(t|X) = \exp(-\beta_0^{(1)} + 0.1B + 0.1W + 0.5W^2 + 0.5X_2)$, and $\lambda_{Z=0}(t|X) = \exp(-0.5 + 0.1B + 0.1W + 0.5W^2 + 0.5X_2)$, the power performance of the test statistics (7) is summarized in Table 12. As expected, the AIPWCC stratified test statistics are more powerful than weighted test statistics. Again, we computed the gold standard test by fitting a proportional hazard Cox model with covariate X and treatment assignment Z on the subsample observations. We use the coefficient of Z divided by its estimated standard deviation as the gold standard test statistic. Note that for those subject not selected in the subsample, their information will not be used to compute the gold standard test statistics. The simulation results indicate that the AIPWCC stratified test statistics are more powerful than the gold standard test. The efficiency gain comes from the use of the main study observations that are not selected in the subsample.

To further illustrate the efficiency gain of the AIPWCC stratified test statistics we also considered a simulation scenario where X_2 is not a confounding variable; that is, the coefficients associated with X_2 in both the regression model and propensity score model were set to zero. Together with the stratified test statistics we also computed the AIPWCC test statistics (3) using only the data from the main study, (AIPWCC (main)) which is a valid test under this scenario. The results are summarized in Table 13. If in fact the additional covariates are not necessary to adjust for confounding, the AIPWCC stratified test statistics are still able to recover much of the information from the main study subjects resulting in comparable power.

We noticed in our simulation studies that the greatest power was obtained with modest truncation; in our case, truncation at 3.5. The value 3.5 corresponds roughly to the 90th percentile of the observed failure times. Thus truncating at the 90th percentile of the observed failure times seems to be a good compromise to balance the instability of the test statistic in the tail of the distribution with the loss of efficiency in not using all the data.

5 Analysis on ASCERT Data

In this section, we apply the proposed log rank test statistics to data from the ASCERT study. The goal is to test the null hypothesis that there is no significant difference between PCI and CABG using baseline covariate data from the main ASCERT study database augmented with information on coronary anatomy from a subsample of patients in the ASCERT angiographic companion study. The main study consisted of records from 9,800 patients in the ASCERT database who underwent a coronary revascularization procedure (CABG or PCI) at one of 54 hospitals participating in the ASCERT study that agreed to be part of the companion study. Subsequently it was determined that some patients were ineligible for analysis. Consequently, the main study participants consisted of 7,393 eligible patients among the 9,800 patients in the 54 hospitals. For the purpose of this analysis we considered the patients from the 54 hospitals of the companion study to be the main focus of inference. Twenty-eight covariates were used on the full sample which included demographics (e.g., age, sex), risk factors (e.g., body mass index, smoking), symptoms and history of cardiovascular disease (e.g., chest pain, congestive heart failure), and comorbidities (e.g., diabetes). The subsample includes records from approximately 2,000 patients chosen by design (roughly 500 in each of the four strata determined by all combinations of the two treatments and whether or not the patients had two- or three-vessel disease). After consideration of eligibility issue, 1,554 eligible patients remained in the sub-study for analysis. Information collected on the subsample includes features of the patient's coronary anatomy (e.g., left-side dominance) and features of each individual blockage (e.g., lesion length, tortuosity, calcification, degree of stenosis).

We are interested in testing the treatment effect between PCI and CABG up to time point 1 month and 4 years. Figure 1 displays the estimated treatment-specific curves using the AIPWCC estimator on the main study (left panel), the stratified AIPWCC estimator with $h = 0$ (middle panel) and with $h = h^{opt}$ (right panel), respectively. The resulting survival curves are very similar: PCI has better short-term survival probability while CABG has better long-term survival probability. This result suggests that the confounding of additional covariates collected in the substudy is not very influential.

Table 14 shows the log rank test statistics on the main study, weighted test statistics, AIPWCC stratified test statistics and ordinary log rank test statistics truncated at 1 month and 4 years. Consistent with the treatment-specific curves, the log rank test statistics in the main study suggests PCI is significantly better than CABG up to 1 month and CABG is better over 4 years. The weighted test statistics only use data from subsample and hence lack power. Meanwhile, the AIPWCC stratified test statistics are able to recover more information by making use of subjects in the main study. For completeness we also computed the ordinary log rank test. In this example the ordinary logrank test exaggerates the treatment difference due to the fact that it does not adjust for confounding (neither in main study nor substudy).

6 Discussion

In this paper, we use semiparametric theory to derive log rank type statistics in observational survival studies. When the important confounding covariates are captured in the main study, the resulting test statistic is doubly-robust. When the issue of confounding is still in question within the main study, a substudy could be conducted to collect additional potential confounding variables. We proposed AIPWCC stratified test statistics in such a scenario, which maintains the double-robustness property.

Acknowledgments

This research was supported by NIH grant 5 R01 HL118336 02.

References

- Bai X, Tsiatis AA, O'Brien SM. Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics*. 2013; 69:830–839. [PubMed: 24117096]
- Hubbard, AE.; van der Laan, MJ.; Robins, JM. Nonparametric locally efficient estimation of the treatment specific survival distributions with right censored data and covariates in observational studies. In: Halloran, E.; Berry, D., editors. *Statistical Models in Epidemiology: The Environment and Clinical Trials*. New York: Springer; 1999. p. 134–178.
- Rubin DB. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*. 1978; 6:34–58.
- Tsiatis, AA. *Semiparametric Theory and Missing Data*. Springer; New York: 2006.
- Weintraub WS, Grau-Sepulveda MV, Weiss JM, O'Brien SM, Peterson ED, Kolm P, Zhang Z, Klein LW, Shaw RE, McKay C, Ritzenthaler LL, Popma JJ, Messenger JC, Shahian DM, Grover FL, Mayer JE, Shewan CM, Garratt KN, Moussa ID, Dangas GD, Edwards FH. Comparative effectiveness of revascularization strategies. *New England Journal of Medicine*. 2012; 366:1467–1476. [PubMed: 22452338]
- Xie J, Liu C. Adjusted Kaplan-Meier Estimator and Log-rank Test with Inverse Probability of Treatment Weighting for Survival Data. *Statistics in Medicine*. 2005; 24:3089–3110. [PubMed: 16189810]
- Zhang M, Schaubel DE. Contrasting treatment-specific survival using double-robust estimators. *Statistics in Medicine*. 2012; 31:4255–4268. [PubMed: 22807175]

Fig. 1.
Estimated AIPWCC and AIPWCC Stratified Survival Estimator for CABG and PCI.

Table 1

Simulation 1, under the null hypothesis, the performance of test statistics (3) when there is no stratified sampling. Sample size $N=2000$. Correct model specification.

truncation	Pr(stat 1.96)	Pr(stat 1.96)	Pr(stat 1.96)
no	0.028	0.026	0.054
at 3.50	0.026	0.026	0.052
at 3.25	0.026	0.026	0.052
at 3.00	0.026	0.026	0.052
at 2.75	0.026	0.028	0.054
at 2.50	0.028	0.030	0.058

Table 2

Simulation 1, under the null hypothesis, the performance of test statistics (3) when there is no stratified sampling. Sample size $N=2000$. Correct regression models and wrong propensity models.

truncation	$\Pr(\text{stat} \geq 1.96)$	$\Pr(\text{stat} \leq -1.96)$	$\Pr(\text{stat} \geq 1.96)$
no	0.026	0.024	0.050
at 3.50	0.028	0.022	0.050
at 3.25	0.030	0.020	0.050
at 3.00	0.028	0.022	0.050
at 2.75	0.028	0.018	0.046
at 2.50	0.028	0.022	0.050

Table 3

Simulation 1, under the null hypothesis, the performance of bootstrap test statistics $G_{bootstrap}$ when there is no stratified sampling. Sample size $N=2000$. Correct regression models and wrong propensity models.

truncation	Pr(stat 1.96)	Pr(stat -1.96)	Pr(stat 1.96)
no	0.030	0.016	0.046
at 3.50	0.024	0.032	0.056
at 3.25	0.024	0.034	0.058
at 3.00	0.026	0.032	0.058
at 2.75	0.026	0.034	0.060
at 2.50	0.028	0.034	0.062

Table 4

Simulation 1, under the null hypothesis, the performance of test statistics (3) when there is no stratified sampling. Sample size $N=2000$. Correct propensity models and wrong regression models.

truncation	$\Pr(\text{stat} \geq 1.96)$	$\Pr(\text{stat} \leq -1.96)$	$\Pr(\text{stat} \geq 1.96)$
no	0.016	0.010	0.026
at 3.50	0.020	0.010	0.030
at 3.25	0.020	0.010	0.030
at 3.00	0.020	0.010	0.030
at 2.75	0.020	0.008	0.028
at 2.50	0.014	0.008	0.022

Table 5

Simulation 1, under the null hypothesis, the performance of Bootstrap test statistics $G_{bootstrap}$ when there is no stratified sampling. Sample size $N=2000$. Correct propensity models and wrong regression models.

truncation	$\Pr(\text{stat} \leq -1.96)$	$\Pr(\text{stat} \geq 1.96)$	$\Pr(\text{stat} \geq 1.96)$
no	0.028	0.020	0.048
at 3.50	0.032	0.018	0.050
at 3.25	0.032	0.016	0.048
at 3.00	0.034	0.022	0.056
at 2.75	0.032	0.022	0.054
at 2.50	0.036	0.018	0.054

Table 6

Simulation 1, under the alternative hypothesis, the power of test statistics (3) and gold standard statistic (GS) when there is no stratified sampling. Sample size N=2000. Correct model specification.

truncation	Power			
	$\beta_0^{(1)}=0.6$		$\beta_0^{(1)}=0.7$	
	AIPWCC	GS	AIPWCC	GS
no	0.346	0.376	0.884	0.906
at 3.50	0.352	0.374	0.898	0.908
at 3.25	0.350	0.376	0.896	0.902
at 3.00	0.352	0.372	0.896	0.908
at 2.75	0.344	0.368	0.886	0.906
at 2.50	0.350	0.370	0.890	0.908

Table 7

Simulation 2, under the null hypothesis, the performance of AIPWCC stratified test statistics and weighted test statistics with stratified sampling. In each cell, the left column corresponds to AIPWCC stratified test statistics and the right column corresponds to weighted test statistics. Sample size $N=2000$, $n_{11} = n_{12} = n_{01} = n_{02} = 300$. Correct model specification.

truncation	Pr(stat 1.96)	Pr(stat 1.96)	Pr(stat 1.96)
no	0.028—0.026	0.024—0.028	0.052—0.054
at 3.50	0.028—0.026	0.026—0.032	0.054—0.058
at 3.25	0.022—0.026	0.030—0.032	0.052—0.058
at 3.00	0.024—0.022	0.030—0.032	0.054—0.054
at 2.75	0.026—0.020	0.028—0.032	0.054—0.052
at 2.50	0.026—0.022	0.026—0.032	0.052—0.054

Table 8

Simulation 2, under the null hypothesis, the performance of AIPWCC stratified test statistics and weighted test statistics with stratified sampling. In each cell, the left column corresponds to AIPWCC stratified test statistics and the right column corresponds to weighted test statistics. Sample size $N=2000$, $n_{11} = n_{12} = n_{01} = n_{02} = 300$. Correct regression models and wrong propensity models.

truncation	Pr(stat ≥ 1.96)	Pr(stat ≤ -1.96)	Pr(stat ≥ 1.96)
no	0.018—0.010	0.024—0.030	0.042—0.040
at 3.50	0.026—0.012	0.026—0.032	0.052—0.044
at 3.25	0.024—0.008	0.032—0.034	0.056—0.042
at 3.00	0.026—0.006	0.024—0.032	0.050—0.038
at 2.75	0.026—0.006	0.024—0.032	0.050—0.038
at 2.50	0.030—0.008	0.026—0.032	0.056—0.040

Table 9

Simulation 2, under the null hypothesis, the performance of AIPWCC bootstrap stratified test statistics $G_{strat,bootstrap}$ and bootstrap weighted test statistics $G_{strat,bootstrap}^{weighted}$ with stratified sampling. In each cell, the left column corresponds to $G_{strat,bootstrap}$ and the right column corresponds to $G_{strat,bootstrap}^{weighted}$. Sample size $N=2000$, $n_{11} = n_{12} = n_{01} = n_{02} = 300$. Correct regression models and wrong propensity models.

truncation	Pr(stat ≥ 1.96)	Pr(stat ≤ -1.96)	Pr(stat ≥ 1.96)
no	0.018—0.010	0.016—0.034	0.034—0.044
at 3.50	0.028—0.014	0.034—0.034	0.062—0.048
at 3.25	0.032—0.010	0.038—0.032	0.070—0.042
at 3.00	0.028—0.008	0.032—0.038	0.060—0.046
at 2.75	0.026—0.012	0.030—0.036	0.056—0.048
at 2.50	0.034—0.010	0.032—0.030	0.066—0.040

Table 10

Simulation 2, under the null hypothesis, the performance of AIPWCC stratified test statistics and weighted test statistics with stratified sampling. In each cell, the left column corresponds to AIPWCC stratified test statistics and the right column corresponds to weighted test statistics. Sample size $N=2000$, $n_{11} = n_{12} = n_{01} = n_{02} = 300$. Correct propensity models and wrong regression models.

truncation	Pr(stat 1.96)	Pr(stat - 1.96)	Pr(stat 1.96)
no	0.018—0.018	0.014—0.008	0.032—0.026
at 3.50	0.022—0.014	0.012—0.010	0.034—0.024
at 3.25	0.018—0.014	0.012—0.012	0.030—0.026
at 3.00	0.018—0.012	0.010—0.012	0.028—0.024
at 2.75	0.016—0.010	0.010—0.014	0.026—0.024
at 2.50	0.018—0.010	0.014—0.016	0.032—0.026

Table 11

Simulation 2, under the null hypothesis, the performance of AIPWCC bootstrap stratified test statistics $G_{strat,bootstrap}$ and bootstrap weighted test statistics $G_{strat,bootstrap}^{weighted}$ with stratified sampling. In each cell, the left column corresponds to $G_{strat,bootstrap}$ and the right column corresponds to $G_{strat,bootstrap}^{weighted}$. Sample size $N=2000$, $n_{11} = n_{12} = n_{01} = n_{02} = 300$. Correct propensity models and wrong regression models.

truncation	Pr(stat 1.96)	Pr(stat -1.96)	Pr(stat 1.96)
no	0.022—0.020	0.024—0.014	0.046—0.034
at 3.50	0.022—0.018	0.028—0.022	0.050—0.040
at 3.25	0.022—0.018	0.028—0.020	0.050—0.038
at 3.00	0.020—0.014	0.024—0.020	0.044—0.034
at 2.75	0.020—0.014	0.026—0.024	0.046—0.038
at 2.50	0.020—0.018	0.028—0.022	0.048—0.040

Simulation 2, under the alternative hypothesis, the power of AIPWCC stratified test statistics (AIPWCC), weighted test statistics (weighted) and gold standard (GS) statistics with stratified sampling. Sample size $N=2000$, $n_{11} = n_{12} = n_{01} = n_{02} = 300$. Correct models specification.

Table 12

truncation	Power					
	$\beta_0^{(1)}=0.6$			$\beta_0^{(1)}=0.7$		
	AIPWCC	weighted	GS	AIPWCC	weighted	GS
no	0.300	0.210	0.268	0.736	0.694	0.760
at 3.50	0.342	0.226	0.272	0.856	0.706	0.748
at 3.25	0.350	0.236	0.268	0.860	0.714	0.752
at 3.00	0.362	0.236	0.270	0.866	0.712	0.758
at 2.75	0.348	0.226	0.258	0.854	0.706	0.748
at 2.50	0.340	0.234	0.270	0.850	0.714	0.748

Table 13

Simulation 2, under the alternative hypothesis, the power comparison of AIPWCC stratified test statistics (AIPWCC), weighted test statistics (weighted), gold standard (GS) statistics with stratified sampling, and main study AIPWCC test statistics (3) (AIPWCC (main)). Sample size $N=2000$, $n_{11} = n_{12} = n_{01} = n_{02} = 300$.

Correct models specification.

truncation	$\beta_0^{(1)}=0.7$			
	AIPWCC	weighted	GS	AIPWCC (main)
no	0.810	0.766	0.808	0.942
at 3.50	0.934	0.774	0.810	0.938
at 3.25	0.932	0.772	0.806	0.940
at 3.00	0.926	0.770	0.800	0.934
at 2.75	0.926	0.772	0.798	0.928
at 2.50	0.922	0.764	0.792	0.922

Table 14

Log rank tests applied on the ASCERT data, including log rank test statistics on main study, weighted test statistics, AIPWCC stratified test statistics and ordinary log-rank test statistics. All tests are truncated at time point of interest: 1 month and 4 years.

Truncation	Main Study	Weighted	AIPWCC Stratified	Ordinary
1 month	2.87	0.39	2.47	2.31
4 years	−2.34	−0.66	−1.91	−3.97