Research Article

Zihan Lin, Ai Ni, and Bo Lu*

# Matched design for marginal causal effect on restricted mean survival time in observational studies

https://doi.org/10.1515/jci-2022-0035
received May 10, 2022; accepted December 31, 2022

**Abstract:** Investigating the causal relationship between exposure and time-to-event outcome is an important topic in biomedical research. Previous literature has discussed the potential issues of using hazard ratio (HR) as the marginal causal effect measure due to noncollapsibility. In this article, we advocate using restricted mean survival time (RMST) difference as a marginal causal effect measure, which is collapsible and has a simple interpretation as the difference of area under survival curves over a certain time horizon. To address both measured and unmeasured confounding, a matched design with sensitivity analysis is proposed. Matching is used to pair similar treated and untreated subjects together, which is generally more robust than outcome modeling due to potential misspecifications. Our propensity score matched RMST difference estimator is shown to be asymptotically unbiased, and the corresponding variance estimator is calculated by accounting for the correlation due to matching. Simulation studies also demonstrate that our method has adequate empirical performance and outperforms several competing methods used in practice. To assess the impact of unmeasured confounding, we develop a sensitivity analysis strategy by adapting the $E$-value approach to matched data. We apply the proposed method to the Atherosclerosis Risk in Communities Study (ARIC) to examine the causal effect of smoking on stroke-free survival.

**Keywords:** confounding bias, marginal effect, noncollapsibility, propensity score matching, sensitivity analysis

**MSC 2020:** 62N02

## 1 Introduction

### 1.1 Causal inference for observational survival data

In biomedical studies, time-to-event is a commonly used outcome measure, and the statistical analyses for such data are usually referred to as survival analysis. Investigating the causal relationship between exposure and the time-to-event outcome is an important topic, with either randomized trials or observational studies. Causal inference for observational survival data has several challenges. First, since not all subjects can be observed for the full duration of time to event, the survival data suffer from censoring, which is a

* **Corresponding author: Bo Lu,** Division of Biostatistics, College of Public Health, The Ohio State University, 244 Cunz Hall, 1841 Neil Avenue, Columbus, OH 43210, United States, e-mail: lu.232@osu.edu
**Zihan Lin:** Division of Biostatistics, College of Public Health, The Ohio State University, 244 Cunz Hall, 1841 Neil Avenue, Columbus, OH 43210, United States, e-mail: lin.2758@osu.edu
**Ai Ni:** Division of Biostatistics, College of Public Health, The Ohio State University, 244 Cunz Hall, 1841 Neil Avenue, Columbus, OH 43210, United States, e-mail: ni.304@osu.edu

type of missing data problem. Therefore, standard statistical methods are usually not sufficient to handle both censoring and the missingness of potential outcomes. Second, the hazard ratio is a popular choice for measuring the association of survival outcomes between two groups, for convenience and easy interpretation. However, the hazard ratio is generally not an appropriate marginal causal effect measure due to its noncollapsibility property [1–3]. Other effect measures need to be considered to warrant valid marginal causal interpretation for survival data. Third, confounding is a major challenge in observational studies, which includes both measured and unmeasured confounders. Propensity score adjustments are popular tools for controlling the observed confounding [4]. But even with successful adjustment of observed confounding, observational data are still vulnerable to unmeasured confounding. Thus, appropriate sensitivity analysis needs to be developed to assess the impact of hidden bias [5].

The issues of using the hazard ratio as a marginal causal effect measure have been discussed extensively in the literature. Greenland et al. [1] pointed out that the hazard ratio has the noncollapsibility property when the treatment effect is nonzero. Hernán [6] argued that using the hazard ratio as a treatment effect measure may not have valid causal interpretation even in randomized studies, since the hazard ratio has a built-in selection bias and may change over time. Martinussen and Vansteelandt [2] studied the estimation of the treatment effect in the presence of confounders and found that the amount of confounding due to noncollapsibility in the Cox proportional hazards (PH) model would be very difficult to quantify. Aalen et al. [7] offered a more theoretical perspective on the conditions under which the hazard has a valid causal interpretation. They suggested that the hazard function $h(t, x, z)$ must satisfy an additive assumption $h(t, x, z) = a(t, z) + b(t, x)$ to yield a causal interpretation, where $a(t, z)$ is a function of survival time $t$ and treatment assignment $z$ and $b(t, x)$ is a function of survival time $t$ and covariates $x$. Ni et al. [8] further illustrated that even under a PH model, the marginal hazard ratio is not a constant, after integrating out covariates. Thus, a valid and simple-to-use causal effect measure for survival outcomes is highly desirable.

## 1.2 RMST difference as a marginal causal effect measure

The restricted mean survival time (RMST) has been used in randomized clinical studies to evaluate treatment effects [9,10]. The RMST difference is more advantageous than the hazard ratio as a marginal effect measure. First, the RMST has an intuitive interpretation as the area under the survival curve over a certain time horizon. Second, the RMST difference is the difference of truncated mean survival time between two groups, which is essentially a mean difference. So it is collapsible, meaning that the marginal and conditional effects are compatible. Third, the treatment effect measured by the RMST difference can be asymptotically unbiasedly estimated without PH assumption, while the conventional Cox model heavily relies on such assumption.

To take advantage of the collapsibility of RMST difference, we can construct RMST regression by including covariates to control for confounding or increase estimation efficiency. Several methods of regressing RMST on multiple covariates have been developed. Karrison [11] examined the RMST as an index for comparing survival in two groups and proposed to model the hazard with piece-wise exponential models assuming covariates have a multiplicative effect on the hazard. Zucker [12] further simplified the implementation procedure for Karrison's method and provided an extended version to achieve robustness against model misspecification. Andersen et al. [13] compared several regression analysis methods of mean survival time and RMST, and they proposed a regression method based on pseudo-observations. Tian et al. [14] developed an RMST regression model with adjustment for baseline covariates. They constructed an estimating equation with the inverse probability of censoring weighting (IPCW) to obtain consistent estimates. Wang and Schaubel [15] modeled the RMST using generalized estimating equation methods, which allows censoring to depend on both baseline covariates and time-dependent factors.

Although RMST differences have been reported in many randomized clinical studies, there is only limited discussion of using RMST in observational studies, probably due to the challenge of confounding adjustment. Propensity score weighting and stratification methods have been explored in the literature. Zhang and Schaubel [16] derived a double-robust estimator for RMST difference based on the inverse

probability of treatment weighted (IPTW) estimating equation with augmentation term. To adjust for confounding factors, they built three working models for survival time, treatment assignment, and censoring, then incorporated them into the augmentation term. They assumed the PH assumption in outcome modeling, which might be violated in practice. Conner et al. [17] proposed a weighted method to compare the adjusted RMST difference directly. Unlike Zhang and Schaubel's work, Conner et al. estimated the RMST based on the Kaplan–Meier (KM) estimator rather than the Nelson–Aalen estimator. They adjusted the KM estimator with IPTW and derived the adjusted RMST by integrating the IPTW-adjusted KM estimator. Ni et al. [8] proposed a propensity score stratified RMST difference estimation strategy to examine the marginal causal effect with observational survival data, which can combine stratification with further regression adjustment. Some existing methods still rely on modeling assumptions, and no matching based method has been proposed. More importantly, none of them touches on unmeasured confounding assessment, which is a big issue in observational studies. The matched design provides more flexibility in confounding adjustment. Observed confounding can be controlled via matching and additional regression modeling. Unobserved confounding can be explored via sensitivity analysis in matched datasets. To address the void of literature, we propose a matching-based RMST difference estimation strategy, which also facilitates the implementation of sensitivity analysis for hidden bias.

## 1.3 A motivating example: atherosclerosis risk in communities (ARIC)

In the United States, stroke is a severe disease that causes serious disability for adults and is a leading cause of death [18,19]. Several previous studies have shown that smoking is an important risk factor for stroke [20,21], and even passive smoking could increase the risk of stroke [22]. Although the causal pathway between smoking and stroke is unclear, Shah and Cole [23] found that the more people smoke, the more likely they were to have a stroke, and people who quit smoking showed a significantly lower risk of stroke, which provides some evidence for the causal relationship between smoking and stroke.

The ARIC study [24] is a prospective cohort study conducted in four U.S. communities. Four thousand adults aged 45–64 years were randomly sampled from each of four U.S. communities, and the final dataset contains information of 15,792 individuals. After a baseline examination during 1987–1989, subjects were followed up for the development of incident ischemic stroke, and the first definite or probable hospitalized stroke. Due to the length of follow-up, not all event times were observed, so the data were subjected to censoring. One primary outcome is the time to first stroke or death (whichever comes first), and a subject is censored if the incidence of stroke or death is not observed by the end of the study. We try to answer a causal question, using this ARIC dataset: how smokers' stroke-free survival would change had they not smoked at baseline. Matched design is a natural choice to address this as it is about the causal effect of those being exposed to smoking, rather than for the entire population.

Existing literature mostly used the Cox PH model to analyze the ARIC data. Kwon et al. [25] and Ding et al. [26] studied the association between smoking status and risk of stroke, using the Cox PH model to estimate the hazard ratio (HR) of smoking status on the risk of stroke. Thus, the estimated effects were interpreted as conditional rather than marginal. Moreover, it is possible that some prognostic factors were not included in the confounder adjusted regression model, which would lead to biased estimates of conditional effects. An analysis with RMST as the effect measure may provide new insight into this research.

In this article, we propose a propensity score matching-based RMST difference estimator and develop a corresponding sensitivity analysis strategy for assessing the impact due to unmeasured confounding. We apply this method to the ARIC study to examine the causal effect of smoking on stroke-free survival. The rest of this article is organized as follows: In Section 2, we set up the notation and assumptions and describe the proposed RMST estimator with its theoretical properties. In Section 3, we conduct a simulation study to examine the empirical performance of our proposed method under different scenarios and also compare it with several commonly used methods in practice. In Section 4, we develop a sensitivity analysis strategy by

adapting the $E$-value approach to matched data. In Section 5, we present the analysis results of the ARIC data. Section 6 concludes this article with some discussions.

# 2 Method: matched RMST difference estimation

## 2.1 Notation and assumptions

We follow the potential outcomes framework proposed by Rubin [27] to define the causal effects. In a two-arm survival analysis study, let $A$ be the treatment assignment indicator (or more generally, the exposure status), such that $A = 1$ indicates being exposed to the treatment and $A = 0$ indicates being exposed to the control. Let $T^a$ denote the potential event time and $S^a(t)$ denote the corresponding survival function for a subject if under treatment value $a$. The following two assumptions are extensions of commonly used assumptions for causal inference in observational studies [28].

**Assumption 1.** Stable unit treatment value assumption (SUTVA). The potential survival times for one individual in the population do not vary with the treatment assigned to others. There are no different versions of the specified treatment level.

**Assumption 2.** Treatment assignment is strongly ignorable given covariates $X$, that is, $(T^0, T^1) \perp\!\!\!\perp A|X$ and $0 < pr(A = 1|X) < 1$.

The potential restricted event time is defined as $Z^a = \min(T^a, \tau)$, where $\tau$ is the truncation time point, which is usually prespecified at the design stage based on clinical relevance and study feasibility. Both $T^a$ and $Z^a$ are subject to censoring by a random variable $C$. We introduce two additional assumptions for survival data.

**Assumption 3.** Censoring is independent of potential survival times and baseline covariates within each treatment group, that is, $C \perp\!\!\!\perp (T^0, T^1)|A$ and $C \perp\!\!\!\perp X|A$.

This assumption ensures that we can asymptotically unbiasedly estimate the survival function via the KM approach within the matched sample. This also implies the conditional independence between censoring and the truncated survival times, $Z^a$.

**Assumption 4.** The truncation time point is smaller than the largest follow-up time, $\tau < t_{\max}$, where $t_{\max}$ is the largest follow-up time (event or censored).

Assumption 4 is a technical one to ensure that the prespecified $\tau$ is clinically meaningful, and RMST can be asymptotically unbiasedly estimated.

Let $\delta^a = I(Z^a < C)$ denote the censoring indicator, and then the observed restricted time is defined as $Y^a = \min(Z^a, C) = (Z^a)^{\delta^a} C^{(1-\delta^a)}$. For a subject under treatment value $a$, the potential outcome of RMST is defined as $\mu^a(\tau) = E(Z^a) = \int_0^\tau S^a(t)\mathrm{d}t$, then the average treatment effect (ATE) on RMST, denoted by $\Delta_{ATE}$, can be defined as follows:

$$\Delta_{ATE} = \mu^1(\tau) - \mu^0(\tau) = E(Z^1) - E(Z^0) = \int_0^\tau [S^1(t) - S^0(t)]\mathrm{d}t.$$

Then the average treatment effect for the treated (ATT) on RMST, denoted by $\Delta_{\mathrm{ATT}}$, can be defined as follows:

$$\Delta_{\text{ATT}} = \mu^1_{A=1}(\tau) - \mu^0_{A=1}(\tau) = E(Z^1|A=1) - E(Z^0|A=1) = \int_0^\tau [S^1_{A=1}(t) - S^0_{A=1}(t)]\mathrm{d}t,$$

which is more meaningful for many observational study applications, where only a portion of the population, not everyone, could have been exposed to the treatment. Since $Z^a = \min(T^a, \tau)$ and $\tau$ is a fixed constant, $(T^0, T^1) \perp\!\!\!\perp A|X$ implies $(Z^0, Z^1) \perp\!\!\!\perp A|X$, and similar conclusions could be made about assumption 3. Following Theorem 3 in Rosenbaum and Rubin [4], we can establish the strong ignorability based on propensity score $e(X) = P(A = 1|X)$ for survival outcomes in proposition 1 (proof provided in Appendix A).

**Proposition 1.** *Given Assumptions* 1 *and* 2, *we have* $(T^0, T^1) \perp\!\!\!\perp A|e(X)$, *which further implies* $(Z^0, Z^1) \perp\!\!\!\perp A|e(X)$.

## 2.2 Matched RMST difference estimator

In randomized trials, the marginal causal effect of the treatment on RMST can be asymptotically unbiasedly estimated [29] by direct contrast of group-specific RMST estimates since confounding effects are eliminated by design. In observational studies, however, additional adjustments are needed for confounding control. Propensity score-based approaches are popular for this purpose, which may take the form of matching, stratification, or weighting [4,30]. Among different propensity score adjustment strategies, matching is a design tool that selects comparable control units to match with treated units, and it often results in more robust causal effect estimates as it does not rely on outcome model specification. Usually, matching uses all treated and a subset of control units, so it estimates the ATT [28].

Our proposed propensity score matched RMST estimation includes the following steps:

(1) *Propensity score estimation*. The propensity score is defined as the conditional probability of treatment given a vector of observed covariates [4]. We estimate the propensity score by fitting a logistic regression on $A$ with $X$, though other estimation options, either parametric or nonparametric, are also available [31, 32].

(2) *Propensity score matching*. We use the optimal matching algorithm by Hansen and Klopfer [33] to create pair matches without replacement based on the estimated propensity score, and the unmatched controls will be removed from the matched sample. Matching quality is assessed by checking the postmatching covariate balance. Any substantial covariate imbalance would lead to a recalibration of the propensity score model. We will proceed to the next step only after a satisfactory balance is achieved. Note that matching is used as a design procedure, and the specific propensity score values are not used in the subsequent estimation process (e.g., not a part of any estimating equations). So the uncertainty of propensity score estimation is not considered in subsequent analysis and variance calculation.

(3) *Treatment effect estimation*. Suppose we obtain $n$ pairs of data through matching, where each pair contains exactly one treated and one control subject. We estimate the RMST, $\mu(\tau)$, by $\hat{\mu}(\tau) = \int_0^\tau \hat{S}(t)\mathrm{d}(t)$, where $\hat{S}(t)$ is estimated by the nonparametric KM method. Let $\hat{S}^0(t)$ and $\hat{S}^1(t)$ denote the KM estimates of survival function for control and treated groups in the matched sample, respectively. On the basis of the matched sample, our estimator for the averaged treatment effect on the treated (ATT) is

$$\hat{\Delta}_{\text{ATT}} = \hat{\mu}^1(\tau) - \hat{\mu}^0(\tau) = \int_0^\tau [\hat{S}^1(t) - \hat{S}^0(t)]\mathrm{d}t.$$

The following two propositions show that the matched RMST difference estimator is asymptotically unbiased (both proofs are provided in Appendix A).

**Proposition 2.** *Given Assumptions* 1–4, *the RMST estimator based on the KM method given propensity score* $e(X)$ *and treatment group* $A$, *denoted as* $\hat{\mu}_{e(X),A}$, *is an asymptotically unbiased estimator for* $\mu_{e(X),A}$ *given* $\tau < t_{\max}$.

**Proposition 3.** *Given Assumptions* 1–4, $\hat{\Delta}_{\mathrm{ATT}}$ *is asymptotically unbiased.*

## 2.3 Variance estimation

The matching process may introduce correlations between the two subjects in the same pair, as they are matched on similar propensity scores. Therefore, the variance calculation of $\hat{\Delta}_{\mathrm{ATT}}$ needs to account for such correlation:

$$\mathrm{var}(\hat{\Delta}_{\mathrm{ATT}}) = \mathrm{var}\left[\int_0^\tau \hat{S}^0(t_0)\mathrm{d}t_0\right] + \mathrm{var}\left[\int_0^\tau \hat{S}^1(t_1)\mathrm{d}t_1\right] - 2\mathrm{cov}\left[\int_0^\tau \hat{S}^0(t_0)\mathrm{d}t_0, \int_0^\tau \hat{S}^1(t_1)\mathrm{d}t_1\right].$$

The overall variance has two components, the marginal variance of RMST estimates and their covariance. For two dependent event times with independent censoring and no competing risk, Murray and Cole [34] provided closed-form asymptotic covariance formulas for KM survival estimates and corresponding RMST estimates. To address the dependence structure introduced in the matching process, we adapt their formulas to compute the covariance between the control and treated group RMST estimates in the matched sample.

Specifically, let $T_0$ be the event time for a subject from the control group with marginal hazard function $h_0(\cdot)$, and $T_1$ be the event time for a subject from the treatment group with marginal hazard function $h_1(\cdot)$, then the event times for a matched pair of control and treated subject can be denoted as $(T_0, T_1)$. Let $C_0$ and $C_1$ be the censoring variables for the control and treated subject, respectively. Then the observed time can be denoted as $\tilde{T}_0 = \min(T_0, C_0)$ for control group with censoring indicator $\delta_0 = I(T_0 < C_0)$ and $\tilde{T}_1 = \min(T_1, C_1)$ for treated group with censoring indicator $\delta_1 = I(T_1 < C_1)$. Then, the joint hazard function is $h_{ij}(u, v) = \lim_{\Delta u, \Delta v \to 0} \frac{1}{\Delta u \Delta v} P(u \le \tilde{T}_i < u + \Delta u, v \le \tilde{T}_j < v + \Delta v, \delta_i = 1, \delta_j = 1 | \tilde{T}_i \ge u, \tilde{T}_j \ge v)$ where $i, j \in \{0, 1\}$, and the conditional hazard function is $h_{i|j}(u|v) = \lim_{\Delta u \to 0} \frac{1}{\Delta u} P(u \le \tilde{T}_i < u + \Delta u, \delta_i = 1 | \tilde{T}_i \ge u, \tilde{T}_j \ge v)$, where $i, j \in \{0, 1\}$. Then, the covariance between two RMSTs can be computed as follows:

$$\begin{aligned}\mathrm{cov}\left[\int_0^\tau \hat{S}^0(t_0)\mathrm{d}t_0, \int_0^\tau \hat{S}^1(t_1)\mathrm{d}t_1\right] &= \frac{1}{n}\int_0^\tau\int_0^\tau \hat{S}^0(t_0)\hat{S}^1(t_1)\int_0^{t_0}\int_0^{t_1} G_{01}(u, v)\mathrm{d}v\mathrm{d}u\mathrm{d}t_0\mathrm{d}t_1 \\ &= \frac{1}{n}\int_0^\tau\int_0^\tau \left[\int_v^\tau \hat{S}^0(t)\mathrm{d}t\right]\left[\int_u^\tau \hat{S}^1(t)\mathrm{d}t\right] G_{01}(u, v)\mathrm{d}v\mathrm{d}u,\end{aligned}$$

where $G_{01}(u, v) = \frac{P(\tilde{T}_0 \ge u, \tilde{T}_1 \ge v)}{P(\tilde{T}_0 \ge u)P(\tilde{T}_1 \ge v)}[h_{01}(u, v) - h_{0|1}(u|v)h_1(v) - h_{1|0}(v|u)h_0(u) + h_0(u)h_1(v)]$. Details about the computation of function $G_{01}(u, v)$ are included in Appendix C.

For the marginal variances, two methods may be considered:

(1) Murray's method: The aforementioned covariance formulas can be used to compute the marginal variance, since the marginal variance of RMST could be written as the covariance with itself, that is, $\mathrm{var}\left(\int_0^\tau \hat{S}(t)\mathrm{d}t\right) = \mathrm{cov}\left[\int_0^\tau \hat{S}(t)\mathrm{d}t, \int_0^\tau \hat{S}(t)\mathrm{d}t\right]$.

(2) Hosmer's method: we may also consider the computation method introduced in the study by Hosmer et al. [35]. Let $t_1 < t_2 < \cdots < t_D$ represent distinct event times. For each $k = 1, \ldots, D$, let $Y_k$ be the number of

surviving units just prior to event time $t_k$, and let $d_k$ be the number of events at $t_k$. Let $\hat{S}(t_k) = \prod_{l=1}^{k}\left(1 - \frac{d_l}{Y_l}\right)$ denotes the KM estimate of the survival function at event time $t_k$, and let $N_\tau$ be the number of $t_k$ values that are less than truncation time point $\tau$, and then the RMST is estimated by

$$\int_0^\tau \hat{S}(t)\mathrm{d}t = \sum_{k=1}^{N_\tau} \hat{S}(t_{k-1})(t_k - t_{k-1}) + \hat{S}\left(t_{N_\tau}\right)\left(\tau - t_{N_\tau}\right),$$

and the marginal variance of RMST can be estimated as follows:

$$\mathrm{var}\left(\int_0^\tau \hat{S}(t)\mathrm{d}t\right) = \frac{m}{m-1}\sum_{k=1}^{N_\tau} \frac{d_k A_k^2}{Y_k(Y_k - d_k)},$$

where $A_k = \int_{t_k}^{\tau} \hat{S}(t)\mathrm{d}t = \sum_{l=k}^{N_\tau} \hat{S}(t_l)(t_{l+1} - t_l) + \hat{S}(t_{N_\tau})(\tau - t_{N_\tau})$ and $m = \sum_{l=1}^{N_\tau} d_l$.

In our simulation studies, we present the variance estimates under Murray's method since the results from these two methods turn out to be very close.

# 3 Simulation studies

## 3.1 Data generation

To assess the empirical performance of the proposed method, we simulate an observational dataset with known confounders. Several existing methods for causal inference with survival outcomes are compared.

We generate ten independent baseline covariates denoted by $X_1$ to $X_{10}$. Among them, $X_1, X_3, \ldots, X_9$ are five binary covariates following Bernoulli distribution with parameters 0.2, 0.4, 0.6, 0.8, and 0.5, respectively, and $X_2, X_4, \ldots, X_{10}$ are five continuous covariates following the standard normal distribution. We then generate potential survival time $T^1$ as the outcome under treatment and potential survival time $T^0$ as the outcome under control from Weibull distribution [36]. Specifically, we simulate a uniform random variable $Q$ on [0,1] and then generate the potential survival time as follows ($j = 1, 0$):

$$T^j = \left(-\frac{\log(Q)}{\lambda_{0j}\exp(\beta_A j + X_1 + 1.2X_4 + 1.4X_6 + 1.6X_7 + 1.6X_8 + 1.4X_9 + 1.2X_{10})}\right)^{\frac{1}{\nu_j}},$$

where $A$ is the treatment indicator and $\beta_A$ is the conditional multiplicative treatment effect on the hazard function given covariates, and $\nu_j$ and $\lambda_{0j}$ are the shape and baseline scale parameters of Weibull distribution for treatment group $j$, respectively. When $\nu_0 = \nu_1$, we have the PH model, otherwise the model is nonproportional hazards (NP). The treatment indicator $A$ is generated from Bernoulli distribution with $P(A = 1|X)$ defined by the logistic model $\mathrm{logit}(P(A = 1|X)) = -1.95 + \log(1.2)X_1 + \log(1.1)X_2 + \log(1.4)X_3 + \log(1.2)X_4 + \log(1.6)X_5 + \log(1.3)X_6 + \log(1.8)X_7$. Thus, $X_1$, $X_4$, $X_6$, and $X_7$ are true confounders. This setup allows about 20% of the population to be exposed to treatment.

In the simulation, we assume censoring variable $C$ is marginally independent of $T^a$ and $X$ over treatment indicator $A$ for simplicity, and $C$ is generated from an exponential distribution with rate parameter $\gamma$, which is chosen to create four different levels of censoring. For simplicity, we use the same censoring variable for both arms in the simulation.

Let $\tau$ be the prespecified truncation time point, and the observed event time is $T = T^0(1 - A) + T^1A$. We generate the restricted event time $Z = \min(T, \tau)$ and the observed restricted time $Y = \min(Z, C) = \min(T, C, \tau)$. The restricted event time $Z$ is censored if the observed time $C < Z$ with censoring status $\delta_Z = I(Z < C)$, otherwise it is noncensored.

We simulate 500 datasets of sample size 2,500 for each scenario and set the truncation time point $\tau$ to 100. The true RMST difference is determined by calculating the empirical difference between the potential RMSTs under treated and control conditions, and we compute both ATT and ATE versions of true RMST difference to serve as benchmarks for different methods as appropriate. In the $j$th simulated dataset, we calculate $\Delta_j = \sum_{i=1}^{n} \frac{Z_i^1 - Z_i^0}{n}$, where $Z_i^A = \min(T_i^A, \tau)$ is the potential restricted event time for the $i$th individual and $n$ is the sample size of the treated group (for ATT) or the entire sample (for ATE). Then, the true marginal effect on RMST is calculated as $\Delta_0 = \sum_{j=1}^{500} \frac{\Delta_j}{500}$.

Both PH and NP settings are examined. Under both settings, we set $\beta_A$ to five different values: 0, −0.4, −0.8, −1.2, and −2. For each treatment effect value, we also consider four different levels of censoring rates (CRs), which are 0, 20, 40, and 60%. Detailed parameter setup for PH and NP scenarios in observational studies are summarized in Table 1.

## 3.2 Estimation strategies

The proposed method is compared with three existing estimation strategies:

(1) *Propensity score matched RMST estimation*. This is our proposed method as described in the previous section, and the propensity score is estimated using the correct model specification. The estimated treatment effect is compared to the ATT version of the true RMST difference in our simulation.

(2) *Conner's IPTW RMST estimation*. This method is proposed by Conner et al. [17], and they estimated the RMST based on the inverse probability treatment weighting (IPTW) adjusted KM estimator. In our simulation, we use the ATT version of weight to adjust for observed confounding, so it is compared to the true ATT RMST difference. The propensity score is estimated using the correct model specification.

(3) *Tian's RMST regression*. This method is proposed by Tian et al. [14], which uses the IPCW estimating equation with identity link function to estimate the treatment effect on RMST with adjustment for covariates. The estimated treatment effect is compared to ATE version of the true RMST difference. We consider four different outcome models in Tian's RMST regression: (1) outcome model using the treatment indicator only; (2) outcome model using the true covariate set; (3) outcome model using all covariates; and (4) outcome model using a wrong covariate set. Due to space limitation, only the results of RMST regression with true covariates are summarized in the following section, which has the best performance among the four models. An important caveat is that the RMST regression model with the true covariate set does not represent the true outcome model since the data are generated based on a hazard model.

**Table 1:** Simulation studies: parameter setup for observational studies scenarios with independent censoring

| $(\nu_0, \lambda_0)$ | $(\nu_1, \lambda_1)$ | $\beta_A$ | Rate parameter gamma | | | |
|---|---|---|---|---|---|---|
| | | | 0% | 20% | 40% | 60% |
| **Nonrandomized PH** | | | | | | |
| (1, exp(−6)) | (1, exp(−6)) | 0 | $1.00 \times 10^{-8}$ | 0.0051 | 0.0142 | 0.0467 |
| (1, exp(−6)) | (1, exp(−6)) | −0.4 | $1.00 \times 10^{-8}$ | 0.004616 | 0.0124 | 0.0345 |
| (1, exp(−6)) | (1, exp(−6)) | −0.8 | $1.00 \times 10^{-8}$ | 0.00421 | 0.011 | 0.0272 |
| (1, exp(−6)) | (1, exp(−6)) | −1.2 | $1.00 \times 10^{-8}$ | 0.003872 | 0.00992 | 0.0226 |
| (1, exp(−6)) | (1, exp(−6)) | −2 | $1.00 \times 10^{-8}$ | 0.0034 | 0.0084 | 0.01731 |
| **Nonrandomized nonPH** | | | | | | |
| (1, exp(−6)) | (1, exp(−6)) | 0 | $1.00 \times 10^{-8}$ | 0.0051 | 0.0142 | 0.0467 |
| (1, exp(−6)) | $(1.5, 1.23 \times 10^{-4})$ | −0.4 | $1.00 \times 10^{-8}$ | 0.003644 | 0.00904 | 0.0189 |
| (1, exp(−6)) | $(1.5, 1.23 \times 10^{-4})$ | −0.8 | $1.00 \times 10^{-8}$ | 0.00343 | 0.0084 | 0.01692 |
| (1, exp(−6)) | $(1.5, 1.23 \times 10^{-4})$ | −1.2 | $1.00 \times 10^{-8}$ | 0.00324 | 0.00787 | 0.01542 |
| (1, exp(−6)) | $(1.5, 1.23 \times 10^{-4})$ | −2 | $1.00 \times 10^{-8}$ | 0.00295 | 0.00702 | 0.01335 |

**Table 2:** Simulation studies: results for proportional hazards scenarios with independent censoring

| Scenario | | Bias (%) | CP | SEM | SEE | Bias (%) | CP | SEM | SEE | Bias (%) | CP | SEM | SEE | Bias (%) | CP | SEM | SEE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\beta_A$ | CR | Matched RMST (Murray) | | | | Tian's RMST regression | | | | Conner's IPTW RMST | | | | IPTW Cox (HR) | | | |
| 0 | 0 | 0.032 | 0.964 | 2.795 | 2.611 | −0.020 | 0.932 | 1.322 | 1.402 | 0.076 | 0.962 | 2.284 | 2.127 | 0.005 | 0.938 | 0.052 | 0.052 |
| | 0.2 | 0.137 | 0.958 | 2.881 | 2.666 | 0.016 | 0.938 | 1.481 | 1.538 | 0.154 | 0.960 | 2.353 | 2.197 | 0.005 | 0.938 | 0.067 | 0.069 |
| | 0.4 | 0.190 | 0.956 | 3.066 | 2.903 | 0.007 | 0.928 | 1.924 | 2.030 | 0.211 | 0.962 | 2.501 | 2.424 | 0.009 | 0.928 | 0.074 | 0.078 |
| | 0.6 | 0.343 | 0.961 | 4.185 | 4.115 | 0.476 | 0.800 | 4.735 | 7.362 | 0.360 | 0.954 | 3.509 | 3.448 | 0.011 | 0.937 | 0.085 | 0.087 |
| −0.4 | 0 | 0.724% | 0.970 | 2.795 | 2.588 | 2.746% | 0.944 | 1.326 | 1.369 | 1.581% | 0.958 | 2.284 | 2.101 | −1.062% | 0.944 | 0.053 | 0.052 |
| | 0.2 | 2.740% | 0.964 | 2.873 | 2.643 | 3.399% | 0.940 | 1.479 | 1.510 | 3.184% | 0.956 | 2.347 | 2.173 | −1.122% | 0.938 | 0.069 | 0.072 |
| | 0.4 | 4.695% | 0.966 | 3.027 | 2.807 | 3.944% | 0.936 | 1.862 | 1.965 | 4.833% | 0.960 | 2.471 | 2.327 | −1.533% | 0.946 | 0.076 | 0.077 |
| | 0.6 | 5.925% | 0.960 | 3.686 | 3.572 | 7.094% | 0.874 | 4.089 | 4.811 | 9.209% | 0.964 | 3.022 | 2.940 | −2.022% | 0.936 | 0.086 | 0.087 |
| −0.8 | 0 | 0.364% | 0.970 | 2.785 | 2.570 | 3.819% | 0.938 | 1.329 | 1.330 | 0.794% | 0.962 | 2.271 | 2.078 | −0.448% | 0.944 | 0.055 | 0.053 |
| | 0.2 | 1.255% | 0.964 | 2.857 | 2.635 | 4.190% | 0.938 | 1.478 | 1.482 | 1.479% | 0.962 | 2.328 | 2.151 | −0.475% | 0.938 | 0.072 | 0.074 |
| | 0.4 | 2.193% | 0.962 | 2.987 | 2.734 | 4.577% | 0.938 | 1.817 | 1.830 | 2.237% | 0.968 | 2.434 | 2.262 | −0.528% | 0.944 | 0.080 | 0.080 |
| | 0.6 | 1.638% | 0.964 | 3.416 | 3.260 | 5.107% | 0.918 | 3.316 | 3.609 | 3.175% | 0.958 | 2.788 | 2.672 | −0.755% | 0.936 | 0.089 | 0.089 |
| −1.2 | 0 | 0.144% | 0.966 | 2.765 | 2.550 | 4.838% | 0.934 | 1.333 | 1.304 | 0.481% | 0.958 | 2.245 | 2.072 | −0.260% | 0.948 | 0.058 | 0.057 |
| | 0.2 | 0.706% | 0.960 | 2.831 | 2.607 | 4.991% | 0.944 | 1.479 | 1.431 | 0.901% | 0.964 | 2.298 | 2.139 | −0.227% | 0.932 | 0.077 | 0.078 |
| | 0.4 | 1.476% | 0.966 | 2.944 | 2.649 | 5.634% | 0.936 | 1.788 | 1.762 | 1.438% | 0.968 | 2.390 | 2.197 | −0.260% | 0.946 | 0.084 | 0.083 |
| | 0.6 | 1.140% | 0.964 | 3.256 | 3.054 | 4.877% | 0.944 | 2.895 | 2.940 | 1.878% | 0.956 | 2.644 | 2.539 | −0.391% | 0.942 | 0.093 | 0.093 |
| −2 | 0 | 0.077% | 0.964 | 2.698 | 2.531 | 6.604% | 0.854 | 1.338 | 1.308 | 0.296% | 0.960 | 2.160 | 2.042 | −0.094% | 0.958 | 0.067 | 0.065 |
| | 0.2 | 0.409% | 0.958 | 2.755 | 2.545 | 6.654% | 0.876 | 1.481 | 1.474 | 0.571% | 0.954 | 2.206 | 2.087 | 0.066% | 0.956 | 0.089 | 0.087 |
| | 0.4 | 0.598% | 0.968 | 2.848 | 2.635 | 6.796% | 0.890 | 1.754 | 1.785 | 0.684% | 0.952 | 2.280 | 2.153 | −0.058% | 0.954 | 0.097 | 0.096 |
| | 0.6 | 0.551% | 0.968 | 3.044 | 2.828 | 6.250% | 0.908 | 2.484 | 2.575 | 0.758% | 0.970 | 2.440 | 2.295 | −0.228% | 0.958 | 0.106 | 0.104 |

Under zero treatment effect scenarios, bias is reported instead of percentage bias.

**Table 3:** Simulation studies: results for nonproportional hazards scenarios with independent censoring

| Scenario | | Bias (%) | CP | SEM | SEE | Bias (%) | CP | SEM | SEE | Bias (%) | CP | SEM | SEE | Bias (%) | CP | SEM | SEE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\beta_A$ | CR | Matched RMST (Murray) | | | | Tian's RMST regression | | | | Conner's IPTW RMST | | | | IPTW Cox (HR) | | | |
| 0 | 0 | 0.032 | 0.964 | 2.795 | 2.611 | −0.020 | 0.932 | 1.322 | 1.402 | 0.076 | 0.962 | 2.284 | 2.127 | 0.005 | 0.938 | 0.052 | 0.052 |
| | 0.2 | 0.137 | 0.958 | 2.881 | 2.666 | 0.016 | 0.938 | 1.481 | 1.538 | 0.154 | 0.960 | 2.353 | 2.197 | 0.005 | 0.938 | 0.067 | 0.069 |
| | 0.4 | 0.190 | 0.956 | 3.066 | 2.903 | 0.007 | 0.928 | 1.924 | 2.030 | 0.211 | 0.962 | 2.501 | 2.424 | 0.009 | 0.928 | 0.074 | 0.078 |
| | 0.6 | 0.343 | 0.961 | 4.185 | 4.115 | 0.476 | 0.800 | 4.735 | 7.362 | 0.360 | 0.954 | 3.509 | 3.448 | 0.011 | 0.937 | 0.085 | 0.087 |
| −0.4 | 0 | 0.196% | 0.964 | 2.674 | 2.473 | 6.560% | 0.888 | 1.244 | 1.215 | 0.428% | 0.964 | 2.131 | 1.966 | 102.646% | 0.000 | 0.056 | 0.061 |
| | 0.2 | 0.684% | 0.966 | 2.745 | 2.502 | 6.777% | 0.894 | 1.386 | 1.347 | 0.835% | 0.964 | 2.192 | 2.017 | 299.268% | 0.000 | 0.079 | 0.081 |
| | 0.4 | 1.150% | 0.964 | 2.860 | 2.583 | 7.132% | 0.902 | 1.659 | 1.643 | 1.133% | 0.964 | 2.292 | 2.106 | 363.043% | 0.000 | 0.091 | 0.094 |
| | 0.6 | 1.168% | 0.972 | 3.122 | 2.917 | 6.584% | 0.920 | 2.427 | 2.479 | 1.389% | 0.966 | 2.521 | 2.361 | 420.641% | 0.000 | 0.105 | 0.109 |
| −0.8 | 0 | 0.158% | 0.966 | 2.644 | 2.461 | 7.174% | 0.820 | 1.253 | 1.212 | 0.345% | 0.958 | 2.091 | 1.954 | 33.373% | 0.004 | 0.057 | 0.062 |
| | 0.2 | 0.425% | 0.968 | 2.710 | 2.468 | 7.298% | 0.850 | 1.394 | 1.351 | 0.561% | 0.966 | 2.148 | 1.991 | 139.830% | 0.000 | 0.083 | 0.085 |
| | 0.4 | 0.801% | 0.968 | 2.815 | 2.569 | 7.541% | 0.876 | 1.656 | 1.660 | 0.845% | 0.964 | 2.240 | 2.072 | 172.413% | 0.000 | 0.096 | 0.099 |
| | 0.6 | 0.906% | 0.966 | 3.032 | 2.831 | 7.119% | 0.890 | 2.316 | 2.411 | 1.014% | 0.968 | 2.429 | 2.308 | 200.416% | 0.000 | 0.110 | 0.114 |
| −1.2 | 0 | 0.079% | 0.966 | 2.605 | 2.434 | 7.806% | 0.724 | 1.261 | 1.224 | 0.263% | 0.960 | 2.042 | 1.925 | 10.204% | 0.446 | 0.059 | 0.063 |
| | 0.2 | 0.321% | 0.962 | 2.668 | 2.444 | 7.912% | 0.772 | 1.402 | 1.394 | 0.463% | 0.966 | 2.095 | 1.960 | 87.060% | 0.000 | 0.088 | 0.091 |
| | 0.4 | 0.565% | 0.970 | 2.764 | 2.524 | 7.969% | 0.828 | 1.652 | 1.661 | 0.626% | 0.956 | 2.179 | 2.027 | 109.224% | 0.000 | 0.103 | 0.106 |
| | 0.6 | 0.599% | 0.980 | 2.950 | 2.755 | 7.634% | 0.884 | 2.243 | 2.280 | 0.652% | 0.958 | 2.340 | 2.242 | 127.609% | 0.000 | 0.117 | 0.121 |
| −2 | 0 | 0.053% | 0.968 | 2.514 | 2.362 | 9.169% | 0.512 | 1.279 | 1.238 | 0.204% | 0.958 | 1.923 | 1.827 | -8.272% | 0.274 | 0.063 | 0.067 |
| | 0.2 | 0.143% | 0.962 | 2.569 | 2.374 | 9.204% | 0.590 | 1.419 | 1.426 | 0.287% | 0.952 | 1.969 | 1.864 | 45.545% | 0.000 | 0.101 | 0.103 |
| | 0.4 | 0.253% | 0.966 | 2.650 | 2.453 | 9.131% | 0.680 | 1.654 | 1.711 | 0.341% | 0.952 | 2.039 | 1.940 | 59.243% | 0.000 | 0.117 | 0.122 |
| | 0.6 | 0.295% | 0.966 | 2.795 | 2.640 | 8.792% | 0.804 | 2.158 | 2.106 | 0.334% | 0.946 | 2.164 | 2.110 | 70.309% | 0.000 | 0.133 | 0.138 |

Under zero treatment effect scenarios, bias is reported instead of percentage bias.

(4) *Inverse probability treatment weighting* (*IPTW*) *Cox regression*. This method estimates $\beta_A$. The propensity score is estimated using the correct model specification. We use the ATT weight to fit a weighted Cox regression model and regard $\beta_A$ as the truth to calculate the bias and coverage probabilities since there is no single value true marginal hazard ratio. We consider four different outcome models in the IPTW Cox regression : (1) outcome model using the treatment indicator only; (2) outcome model using the true covariate set; (3) outcome model using all covariates; (4) outcome model using a wrong covariate set. Due to space limitation, only the results of IPTW Cox regression with the true covariate set are summarized in the following section, which has the best performance among the four models. We understand that the results here are not directly comparable to the first three methods, as they are based on different effect measures. Due to the high popularity of the IPTW Cox model in practice, however, we think there is some value in presenting the results as a reference.

### 3.3 Performance assessment

We summarize treatment effect estimates from 500 Monte Carlo iterations into four measures: (1) percentage bias (Bias %), which is the bias divided by the true value for nonzero treatment effect scenarios. For the zero treatment effect scenario, we just report the bias. The bias is computed as the average of 500 treatment effect estimates minus the truth; (2) coverage probability (CP), which is the proportion of 500 95% confidence (CIs) intervals that cover the truth; (3) model-based standard error (SEM), which is the average of the 500 estimated standard errors from the model-based formula; and (4) empirical standard error (SEE), which is the standard error of the 500 point estimates of the treatment effect.

### 3.4 Results

Simulation results under the PH setting are summarized in Table 2. The proposed matched RMST method generates unbiased estimates of the target parameters under most scenarios, and the coverage probabilities are around 95%. For a small effect size ($\beta_A = -0.4$), the bias is a bit large for a high CR. Conner's method has a similar performance, with moderately larger biases. Averaging across all scenarios, bias from Conner's method is 65% higher than our method. The results of the IPTW Cox model are mostly good since we use the correct outcome model. The CP may be a bit lower than the nominal level, sometimes, which may be due to the underestimated standard error. Tian's RMST regression method shows a relatively large percentage bias and lower CP, especially under scenarios with large treatment effects. This is likely due to the incorrect covariate functional form specification in the model even though we include the right covariate set.

Simulation results under the NP setting are summarized in Table 3. Both our matched RMST method and Conner's method have similar performance (with the latter having more bias) as under the PH setting since these methods do not rely on the PH assumption. Tian's RMST regression method performs somewhat worse, with a bigger bias and much lower than ideal coverage probabilities. Because the PH assumption does not hold here, the IPTW Cox model completely misses the target with large bias and very small coverage probabilities.

## 4 Sensitivity analysis based on matched design

### 4.1 An overview of *E*-value

Propensity score adjustment can only control for observed confounding. Unmeasured confounding is likely to be present in observational studies since researchers have no control over the treatment assignment. Thus, sensitivity analysis is important to assess the impact of hidden bias.

Ding and VanderWeele [37,38] developed a new sensitivity analysis strategy, known as the $E$-value method. It assumes a hypothetical unmeasured confounder, $U$, and provides a lower bound of the strength of association on the risk ratio scale that $U$ would have both the exposure and the outcome, to explain away the observed association. Below is a brief review of the conventional $E$-value method to set the stage for our sensitivity analysis of RMST difference.

Let $A$ denote a binary exposure and $D$ denote a binary outcome, $X$ is a vector of measured confounders, and $U$ is a binary unmeasured confounder with levels $k = 0, 1$. The observed relative risk of exposure $A$ on the outcome $D$ within stratum of $X = x$ is expressed as follows:

$$\mathrm{RR}_{AD|x}^{\mathrm{obs}} = \frac{P(D = 1|A = 1, X = x)}{P(D = 1|A = 0, X = x)}.$$

Then the relative risk of exposure on level $k$ of the unmeasured confounder $U$ within the stratum of $X = x$ is expressed as follows:

$$\mathrm{RR}_{AU,k|x} = \frac{P(U = k|A = 1, X = x)}{P(U = k|A = 0, X = x)}.$$

Since $U$ is not observed, to facilitate the analysis, we take the maximal relative risk of $A$ on $U$ within stratum $X = x$, denoted as $\mathrm{RR}_{AU|x} = \max_k \mathrm{RR}_{AU,k|x}$. Similarly, we can define an upper bound for the relative risk between $U$ and $D$ as $\mathrm{RR}_{UD|x} = \max(\mathrm{RR}_{UD|A=0,x}, \mathrm{RR}_{UD|A=1,x})$, where $\mathrm{RR}_{UD|A,x}$ is an upper bound of the relative risk between $U$ and $D$ in exposed or unexposed group, within stratum $X = x$. If $X$ and $U$ are sufficient to control for all confounding effects, the true causal relative risk is

$$\mathrm{RR}_{AD|x}^{\mathrm{true}} = \frac{\sum_{k=0}^{1} P(D = 1|A = 1, X = x, U = k)P(U = k|X = x)}{\sum_{k=0}^{1} P(D = 1|A = 0, X = x, U = k)P(U = k|X = x)}.$$

The relative risk pair $(\mathrm{RR}_{AU|x}, \mathrm{RR}_{UD|x})$ are used to measure the strength of confounding between the exposure $A$ and the outcome $D$ induced by the confounder $U$ within the stratum of $X = x$. Even though we cannot estimate the true relative risk, its ratio with the observed relative risk is bounded by the following quantity, which is a function of the sensitivity parameters $\mathrm{RR}_{AU|x}$ and $\mathrm{RR}_{UD|x}$:

$$\frac{\mathrm{RR}_{AD|x}^{\mathrm{obs}}}{\mathrm{RR}_{AD|x}^{\mathrm{true}}} \leq \frac{\mathrm{RR}_{AU|x} \times \mathrm{RR}_{UD|x}}{\mathrm{RR}_{AU|x} + \mathrm{RR}_{UD|x} - 1}.$$

For given values of $\mathrm{RR}_{AU|x}$ and $\mathrm{RR}_{UD|x}$, we can identify a range of possible values for the true relative risk. If the range covers one, the observed significant association would be explained away by the presence of unmeasured confounding at the given magnitude.

## 4.2 Sensitivity analysis on RMST difference with matched data

This $E$-value method can be adapted to conduct sensitivity analysis for our RMST difference estimator in matched design. There are a series of propositions to justify the theoretical validity of using the $E$-value for the RMST difference estimator. In the interest of space, we just illustrate the main idea in this subsection and present the propositions and their detailed proofs in Appendix B.

Let $A$ be the treatment indicator and $Z = \min(T, \tau)$ be the RMST outcome, where $T$ is the event time and $\tau$ is the truncation time point. Let $e(X)$ be the propensity score and $U$ be a binary unmeasured confounder with levels $k = 0, 1$. The relative risk of treatment $A$ on level $k$ of the unmeasured confounder $U$ with a given propensity score value $e(X) = e(x)$ is defined as follows:

$$\mathrm{RR}_{AU,k|e(x)} = \frac{P(U = k|A = 1, e(X) = e(x))}{P(U = k|A = 0, e(X) = e(x))}.$$

The maximal relative risk of $A$ on $U$ with $e(X) = e(x)$ is $\mathrm{RR}_{AU|e(x)} = \max_k \mathrm{RR}_{AU,k|e(x)}$. We define the expectations of the RMST outcome $Z$ given $U = u$ and $e(X) = e(x)$ with and without treatment as follows:

$$r_1(u) = E(Z|A = 1, U = u, e(X) = e(x)),$$
$$r_0(u) = E(Z|A = 0, U = u, e(X) = e(x)).$$

Then, the mean ratios of $U$ on $Z$ with and without treatment with $e(X) = e(x)$ are defined as follows:

$$\mathrm{MR}_{UZ|A=1,e(X)=e(x)} = \frac{\max_u r_1(u)}{\min_u r_1(u)}, \quad \mathrm{MR}_{UZ|A=0,e(X)=e(x)} = \frac{\max_u r_0(u)}{\min_u r_0(u)},$$
$$\mathrm{MR}_{UZ|e(X)=e(x)} = \max(\mathrm{MR}_{UZ|A=1,e(X)=e(x)}, \mathrm{MR}_{UZ|A=0,e(X)=e(x)}).$$

As shown in proposition 4 in Appendix B, since both unmeasured confounder parameters $\mathrm{RR}_{AU|e(X)=e(x)}$ and $\mathrm{MR}_{UZ|e(X)=e(x)}$ are no less than 1, we can identify the bounding factor as follows:

$$BF_{U|e(X)=e(x)} = \frac{\mathrm{RR}_{AU|e(X)=e(x)} \times \mathrm{MR}_{UZ|e(X)=e(x)}}{\mathrm{RR}_{AU|e(X)=e(x)} + \mathrm{MR}_{UZ|e(X)=e(x)} - 1}, \tag{1}$$

where $(\mathrm{RR}_{AU|e(X)=e(x)}, \mathrm{MR}_{UZ|e(X)=e(x)})$ are prespecified sensitivity analysis parameters. In theory, one can identify a separate bounding factor for each matched pair and choose the maximal value to facilitate the calculation. But this could be quite cumbersome in practice. Instead, we follow the idea of using $\Gamma$ in the conventional Rosenbaum's sensitivity analysis, which is a prespecified upper bound of the association. Denote $\mathrm{RR}_{AU} = \max_{e(x)}(\mathrm{RR}_{AU|e(X)=e(x)})$ and $\mathrm{MR}_{UZ} = \max_{e(x)}(\mathrm{MR}_{UZ|e(X)=e(x)})$, then the maximal bounding factor can be calculated as follows:

$$BF_U^* = \frac{\mathrm{RR}_{AU} \times \mathrm{MR}_{UZ}}{\mathrm{RR}_{AU} + \mathrm{MR}_{UZ} - 1}. \tag{2}$$

Because $BF_U^*$ is an increasing function of both $\mathrm{RR}_{AU}$ and $\mathrm{MR}_{UZ}$, taking prespecified upper bounds of both associations leads to an upper bound of bounding factors. In practice, one can identify a range of possible values for $(\mathrm{RR}_{AU}, \mathrm{MR}_{UZ})$ and calculate the upper bound of treatment effects for each combination using the following formulas:

Let $\mathrm{ACE}_{AZ}^{\mathrm{true}}$ denote the true average causal effect. When the treatment effect is positive, we have

$$\mathrm{ACE}_{AZ}^{\mathrm{true}} \geq \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right)E(Z|A = 1) - \frac{1}{2}(1 + BF_U^*)E(Z|A = 0). \tag{3}$$

When the treatment effect is negative, we have

$$\mathrm{ACE}_{AZ}^{\mathrm{true}} \leq \frac{1}{2}(1 + BF_U^*)E(Z|A = 1) - \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right)E(Z|A = 0). \tag{4}$$

The sensitivity analysis is generally done by checking the behavior of the 95% CI bounds. Since our real data analysis shows a negative treatment effect, we focus on equation (4) and use a one-sided 95% CI for illustration purposes. Equation (4) implies that the treatment effect estimate should be bounded by a function of the bounding factor and the mean survival times from each treatment group. Denote the right-hand side quantity as right-hand side (RHS). We can calculate the 95% CI for RHS using normal approximation and denote the upper confidence bound as $\mathrm{RHS}_{ub}$. Because $\mathrm{ACE}_{AZ}^{\mathrm{true}} \leq \mathrm{RHS}$, we have $P(\mathrm{ACE}_{AZ}^{\mathrm{true}} \leq \mathrm{RHS}_{ub}) \geq 0.95$. Therefore, we can regard $\mathrm{RHS}_{ub}$ as an upper bound for the upper 95% confidence bound of the treatment effect estimate. If this value is less than zero, we would reject the null hypothesis. Otherwise, we would retain the null.

## 4.3 Interpreting the sensitivity analysis

For an unmeasured confounding with a prespecified magnitude of associations $(\mathrm{RR}_{AU}, \mathrm{MR}_{UZ})$, the bounding factor $BF_U^*$ can be computed by equation (2). For a positive treatment effect based on the observed data, we can compute a lower bound for the lower 95% confidence bound of the treatment effect estimate by

equation (3). A positive lower bound indicates that there is still a positive treatment effect with an unmeasured confounding effect of magnitude ($RR_{AU}$, $MR_{UZ}$). A nonpositive lower bound indicates that the positive treatment effect could be explained away by the unmeasured confounding of magnitude ($RR_{AU}$, $MR_{UZ}$). For a negative treatment effect based on the observed data, we can compute an upper bound for the upper 95% confidence bound of the treatment effect estimate by equation (4) as described earlier, and similar interpretations can be made. A negative upper bound indicates that there is still a negative treatment effect with an unmeasured confounding effect of magnitude ($RR_{AU}$, $MR_{UZ}$). A nonnegative upper bound indicates that the negative treatment effect could be explained away by the unmeasured confounding of magnitude ($RR_{AU}$, $MR_{UZ}$). A detailed numerical example is presented in Section 5.

# 5 Real data example

In this section, we apply our proposed method to the ARIC data [24] to examine the causal effect of baseline smoking on stroke-free survival. Incident ischemic stroke events or death, the primary outcome, are identified through December 31, 2011. After excluding a small portion of subjects with missing values in the variables of interest, the total sample size used in the analysis is 14,549. The event time is defined as the follow-up time (in months) for the first incident stroke or death, whichever comes first, and a subject is censored if neither incident stroke nor death is observed during the study. There are 5,345 events, corresponding to a 63.3% CR. Given the length of follow-up, we choose 240 months as the truncation time $\tau$ for the RMST calculation. Exposure is defined as the smoking status at baseline. There are 3,832 (26.3%) current smokers at baseline. Eight important baseline covariates are included in the propensity score model: race (black, white), gender (male, female), age (44–66 years), body mass index (BMI) (14.2–65.9), diabetes (1 = yes, 0 = no), high-density lipoprotein (HDL) (10–163 mg/dL), low-density lipoprotein (LDL) (0–504.6 mg/dL), and hypertension (1 = yes, 0 = no). Table 4 summarizes these variables by baseline smoking status.

We first fit a logistic regression model on baseline smoking status using the eight covariates to estimate the propensity score. Then, we conduct a 1–1 optimal pair matching without replacement for all subjects, which results in 3,832 pairs and unmatched nonsmokers being removed from the matched sample. The covariates balance is measured by the standardized mean difference, and Figure 1 shows the covariates balance of ARIC data before and after propensity score matching, which indicates our matching achieves very good covariates balance.

For comparison purposes, the analysis results of the proposed method, Tian's RMST regression, and the IPTW Cox regression are all reported in Table 5. All methods show significant evidence of a harmful effect of smoking on the risk of incident ischemic stroke or death. This conclusion agrees with previous findings in the literature. The matched RMST analysis suggests an average reduction of 22.3 stroke-free survival months for baseline smokers had they not smoked at the baseline. Tian's RMST regression method provides similar results as our proposed method in this dataset. The IPTW Cox regression measures the treatment effect on

**Table 4:** Real data example: summary statistics of covariates by baseline smoking status in ARIC study

| | **Non-current smoker (10,717)** | **Current smoker (3,832)** |
|---|---|---|
| Race, *n* (%) of white | 8,161 (76.2%) | 2,730 (71.2%) |
| Gender, *n* (%) of female | 5,957 (55.6%) | 2,003 (52.3%) |
| Age, mean (SD) | 54.4 (5.8) | 53.7 (5.7) |
| BMI, mean (SD) | 28.1 (5.4) | 26.3 (5.0) |
| Diabetes, *n* (%) | 1,044 (9.7%) | 333 (8.7%) |
| HDL (mmol/L), mean (SD) | 52.6 (16.8) | 49.6 (17.3) |
| LDL (mmol/L), mean (SD) | 137.6 (38.9) | 138.6 (40.4) |
| Hypertension, *n* (%) | 3,783 (35.3%) | 1,225 (32.0%) |

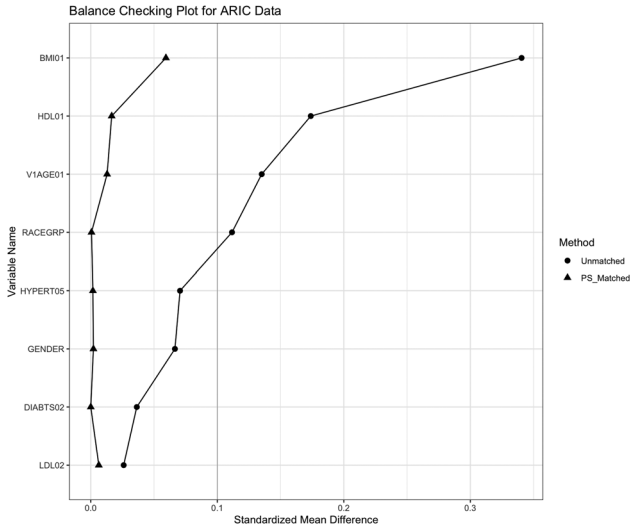


**Figure 1:** Real data example: covariates balance checking.

**Table 5:** Real data example: ARIC data analysis results

| | **Estimate** | **SE** | **95% CI (one-sided)** |
|---|---|---|---|
| Matched RMST | −22.266 | 1.380 | $(-\infty, -19.996]$ |
| Tian's RMST regression | −22.666 | 1.129 | $(-\infty, -20.809]$ |
| IPTW cox (HR) | 2.173 | 0.066 | $[2.068, \infty)$ |

For the matched RMST method and Tian's RMST regression method, the 95% CI bound is the upper bound. For the IPTW Cox method, the 95% CI bound is the lower bound.

the hazard ratio scale, which is not directly comparable to RMST differences. The estimated HR of 2.2 implies that smoking increases the hazard of incident ischemic stroke or death.

All the aforementioned analyses assume the ignorable treatment assignment. However, for such a large observational study, unmeasured confounding is likely to be present, especially given that we are only able

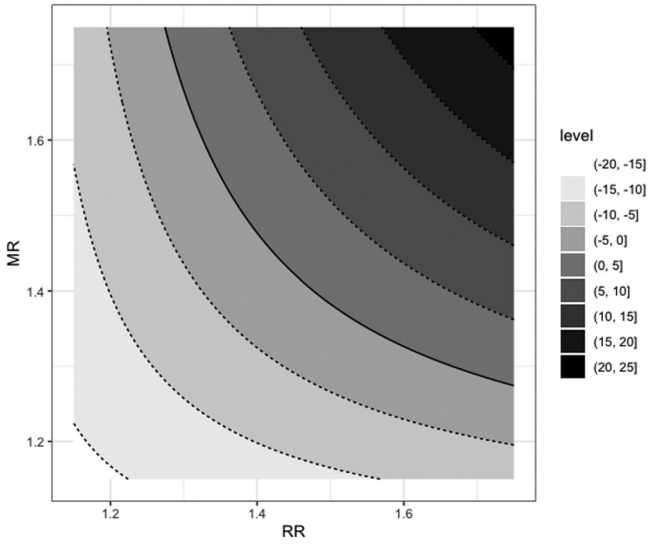


**Figure 2:** Sensitivity analysis: contour plot of 95% CI upper bounds of the treatment effect upper bound. The solid curve represents value 0.

to control a small number of factors. Therefore, it is important to assess how the observed causal effect may change in the presence of hidden bias. A sensitivity analysis, as described in Section 4.2, is carried out for different possible impacts of $U$ on the exposure and the outcome.

We calculate the upper bound of the upper 95% confidence bound of estimated treatment effects by exploring different combinations of $(\mathrm{RR}_{AU}, \mathrm{MR}_{UZ})$, where both values are larger than 1. For illustrative purposes, we focus on a range of values between 1.15 and 1.75 to create a contour plot in Figure 2. Different gray scales reflect different upper bound values of the upper 95% confidence bound of the treatment effect, with the lighter color indicating smaller values and the darker color indicating larger values. The solid curve in the middle of the plot is when the upper bound of the upper 95% confidence bound of the treatment effect is zero.

For the area below this threshold, the upper bound of the upper 95% confidence bound of the treatment effect is still negative, implying a true negative causal effect even in the presence of unmeasured confounding. For the area above this threshold, the upper bound of the upper 95% confidence bound of the treatment effect becomes positive, implying the initial negative treatment effect can be explained away by the presence of unmeasured confounding.

In the real data analysis, we observe a negative treatment effect of −22.266 months with 95% confidence interval (CI) as $(-\infty, -19.996]$ based on our proposed matched RMST method. We also estimate that $E(Z|A = 1) = 195.379$ with SD = 1.087 and $E(Z|A = 0) = 217.646$ with SD = 0.834, and the covariance between these two RMSTs is −0.014. For example, with $\mathrm{RR}_{AU} = 1.4$ and $\mathrm{MR}_{UZ} = 1.45$, we can calculate the bounding factor as follows: $BF_U^* = \frac{1.4 \times 1.45}{1.4 + 1.45 - 1} = 1.097$, then following equation (4), RHS can be calculated as $\mathrm{ACE}_{AZ}^{\mathrm{true}} \le \frac{1}{2}(1 + 1.097) * 195.379 - \frac{1}{2}\left(1 + \frac{1}{1.097}\right) * 217.646 = -3.169 = \mathrm{RHS}$. The variance of RHS is expressed as follows:

$$\begin{aligned}\mathrm{Var(RHS)} = &\left[\frac{1}{2}(1 + 1.097)\right]^2 \cdot 1.087^2 + \left[\frac{1}{2}(1 + \frac{1}{1.097})\right]^2 \cdot 0.834^2 \\ &- 2 \cdot \left[\frac{1}{2}(1 + 1.097)\right] \cdot \left[\frac{1}{2}(1 + \frac{1}{1.097})\right] \cdot (-0.014) = 1.962.\end{aligned}$$

The 95% one-sided CI is $(-\infty, -3.169 + 1.64 * \sqrt{1.962}] = (-\infty, -0.872]$. Since the upper 95% confidence bound is still less than zero, we would reject the null hypothesis of no causal effect. This is robust to unmeasured confounding with the magnitude of impact up to $\mathrm{RR}_{AU} = 1.4$ and $\mathrm{MR}_{UZ} = 1.45$.

For small-to-moderate deviations from the ignorability assumption ($\mathrm{RR}_{AU} < 1.43$ and $\mathrm{MR}_{UZ} < 1.43$), a harmful effect still holds, as the upper bound of the 95% confidence bound of the treatment effect is below the solid zero-curve. For moderate-to-large deviations, the 95% CI upper bound of the treatment effect upper bound may exceed zero, indicating a possibility of a null effect. For example, at (1.5, 1.5), the upper bound of the estimated treatment effect is 2.036 and the upper bound of the 95% confidence bound of the treatment effect is 4.345, which indicates that the harmful treatment effect could be totally explained away by the unmeasured confounding of magnitude $(\mathrm{RR}_{AU}, \mathrm{MR}_{UZ}) = (1.5, 1.5)$. Overall, our sensitivity analysis indicates that the observed significant causal effect is moderately robust to hidden bias.

# 6 Discussion

In this article, we adopt the RMST difference as a marginal causal effect measure for survival data, since it is collapsible and has an easy interpretation. We develop a matching-based RMST difference estimator that is asymptotically unbiased and does not rely on the PH assumption. But this does not rule out the use of hazard in causal analysis with survival data. As pointed out by Aalen et al. [7], the hazard function $h(t, x, z)$ may have a valid causal interpretation, if it satisfies some additive structural constraint.

An interesting practical issue with RMST is the choice of $\tau$. The common practice is to prespecify the truncation time at the study design stage or to make the decision independent of the observed outcomes. It

is usually determined based on content expertise, for example, an important clinical time point for the disease under study. Kim et al. [39] picked a truncation time of 5 years for the Placement of Aortic Transcatheter Valves (PARTNER) trial as they were interested in the effect of transcatheter aortic valve replacement procedure versus routine medical treatment on preventing death in 5 years. Recently, Tian et al. [40] provided a more thorough discussion on the empirical choice of time window in RMST. They also showed that under a mild condition on the censoring distribution, one could make inferences about the RMST up to $\tau$, where $\tau$ could be equal to the largest follow-up time (either observed or censored) in the study. With such choices, RMST incorporates all available information.

One limitation of our work is that the proposed nonparametric estimator may not be easily extended to more complex matching designs, such as 1-$k$ or full matching designs. This is because we need to compute the covariance to account for the correlation in matched sets. But the covariance calculation relies on the assumption of equal sample sizes in both groups [34]. Therefore, the covariance formula can not be applied directly to other matching designs. One strategy to relax this limitation is to consider fitting a parametric RMST regression model after matching. This could be more advantageous if we have a good idea about the outcome model specification, as it may correct residual confounding bias not captured by matching. This adds more flexibility to postmatching inference, as it can lead to more robust or efficient semiparametric strategies by combining matching with regression models [41]. It also makes our method more attractive in practice than Conner's method as the latter solely relies on KM estimation of survival functions and cannot include regression models.

**Acknowledgments:** This work was partially supported by grant DMS-2015552 from National Science Foundation. The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I, and HHSN268201700004I. The authors thank the staff and participants of the ARIC study for their important contributions. The authors also thank the associate editor and two anonymous reviewers for their insightful comments, which lead to substantial improvement of the manuscript.

**Funding information**: This work was partially supported by grant DMS-2015552 from National Science Foundation.

**Author contributions**: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

**Conflict of interest**: The authors state no conflict of interest.

**Informed consent**: Not applicable.

**Ethical approval**: The conducted research uses a de-identified secondary dataset in the analysis, hence not related to either human or animals use.

**Data availability statement**: The ARIC dataset analyzed during the current study is available in the NHLBI Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) [https://biolincc.nhlbi.nih.gov/studies/aric/].

# References

[1] Greenland S, Pearl J, Robins JM. Confounding and collapsibility in causal inference. Stat Sci. 1999;14(1):29–46. doi: 10.1214/ss/1009211805.

[2] Martinussen T, Vansteelandt S. On collapsibility and confounding bias in cox and aalen regression models. Lifetime Data Analysis. 2013;19(3):279–96. doi: 10.1007/s10985-013-9242-z.
[3] Sjooolander A, Dahlqwist E, Zetterqvist J. A note on the noncollapsibility of rate differences and rate ratios. Epidemiology. 2016;27(3):356–9. doi: 10.1097/EDE.0000000000000433.
[4] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. doi: 10.1093/biomet/70.1.41.
[5] Rosenbaum PR. Design of observational studies. 2nd ed. Cham: Springer; 2020. doi: 10.1007/978-3-030-46405-9.
[6] Hernán MA. The hazards of hazard ratios. Epidemiology. 2010;21(1):13. doi: 10.1097/EDE.0b013e3181c1ea43.
[7] Aalen OO, Cook RJ, Røysland K. Does cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Analysis. 2015 June;21(4):579–93. doi: 10.1007/s10985-015-9335-y.
[8] Ni A, Lin Z, Lu B. Stratified restricted mean survival time model for marginal causal effect in observational survival data. Ann Epidemiol. 2021;64:149–54. https://www.sciencedirect.com/science/article/pii/S1047279721003082.
[9] Royston P, Parmar MK. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol. 2013;13(1):1–15. doi: 10.1186/1471-2288-13-152.
[10] Trinquart L, Jacot J, Conner SC, Porcher R. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. J Clin Oncol. 2016;34(15):1813–9. doi: 10.1200/JCO.2015.64.2488.
[11] Karrison T. Restricted mean life with adjustment for covariates. J Am Stat Assoc. 1987;82(400):1169–76. https://www.tandfonline.com/doi/abs/10.1080/01621459.1987.10478555.
[12] Zucker DM. Restricted mean life with covariates: modification and extension of a useful survival analysis method. J Am Stat Assoc. 1998;93(442):702–9. https://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10473722.
[13] Andersen PK, Hansen MG, Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Analysis. 2004;10(4):335–50. doi: 10.1007/s10985-004-4771-0.
[14] Tian L, Zhao L, Wei LJ. Predicting the restricted mean event time with the subjectas baseline covariates in survival analysis. Biostatistics. 2013;15(2):222–33. doi: 10.1093/biostatistics/kxt050.
[15] Wang X, Schaubel DE. Modeling restricted mean survival time under general censoring mechanisms. Lifetime Data Analysis. 2018;24(1):176–99. doi: 10.1007/s10985-017-9391-6.
[16] Zhang M, Schaubel DE. Double-Robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. Biometrics. 2012;68(4):999–1009. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2012.01759.x.
[17] Conner SC, Sullivan LM, Benjamin EJ, LaValley MP, Galea S, Trinquart L. Adjusted restricted mean survival times in observational studies. Stat Med. 2019;38(20):3832–60. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8206.
[18] Kochanek KD, Murphy SL, Xu J, Arias E. Mortality in the United States, 2013. NCHS Data Brief. 2014;(178):1–8.
[19] Mozaffarian D, Benjamin EJ, Go AS, Arnett DK, Blaha MJ, Cushman M, et al.Heart disease and stroke statistics-2016 update. Circulation. 2016;133(4):e38–360. https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000350.
[20] Wolf PA, D'Agostino RB, Kannel WB, Bonita R, Belanger AJ. Cigarette smoking as a risk factor for stroke: the framingham study. JAMA. 1988;259(7):1025–9. doi: 10.1001/jama.1988.03720070025028.
[21] Shinton R, Beevers G. Meta-analysis of relation between cigarette smoking and stroke. British Med J. 1989;298(6676):789–94. https://www.bmj.com/content/298/6676/789.
[22] Bonita R, Duncan J, Truelsen T, Jackson RT, Beaglehole R. Passive smoking as well as active smoking increases the risk of acute stroke. Tobacco Control. 1999;8(2):156–60. https://tobaccocontrol.bmj.com/content/8/2/156.
[23] Shah RS, Cole JW. Smoking and stroke: the more you smoke the more you stroke. Expert Rev Cardiovasc Ther. 2010;8(7):917–32. doi: 10.1586/erc.10.56.
[24] Aric investigators. The atherosclerosis risk in communities (ARIC) study: design and objectives. Am J Epidemiol. 1989;129(4):687–702. doi: 10.1093/oxfordjournals.aje.a115184.
[25] Kwon Y, Norby FL, Jensen PN, Agarwal SK, Soliman EZ, Lip GYH, et al. Association of smoking, alcohol, and obesity with cardiovascular death and ischemic stroke in atrial fibrillation: the atherosclerosis risk in communities (ARIC) study and cardiovascular health study (CHS). Plos One. 2016 Jan;11(1):1–13. doi: 10.1371/journal.pone.0147065.
[26] Ding N, Sang Y, Chen J, Ballew SH, Kalbaugh CA, Salameh MJ, et al. Cigarette smoking, smoking cessation, and long-term risk of 3 major atherosclerotic diseases. J Am College Cardiol. 2019;74(4):498–507. doi: 10.1016/j.jacc.2019.05.049.
[27] Rubin DB. Estimating Causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701. doi: 10.1037/h0037350.
[28] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. New York: Cambridge University Press; 2015. doi: 10.1017/CBO9781139025751.
[29] Fleming TR, Harrington DP. Counting processes and survival analysis. Hoboken, New Jersey: John Wiley & Sons; 2011. doi: 10.1002/9781118150672.
[30] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61(4):962–73. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2005.00377.x.

[31] McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Meth. 2004;9(4):403–25. doi: 10.1037/1082-989X.9.4.403.
[32] Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. J Clin Epidemiol. 2010;63(8):826–33. doi: 10.1016/j.jclinepi.2009.11.020.
[33] Hansen BB, Klopfer SO. Optimal full matching and related designs via network flows. J Comput Graph Stat. 2006;15(3):609–27. doi: 10.1198/106186006X137047.
[34] Murray S, Cole B. Variance and sample size calculations in quality-of-life-adjusted survival analysis (Q-TWiST). Biometrics. 2000;56(1):173–82. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0006-341X.2000.00173.x.
[35] Hosmer DW, Lemeshow S, May S. Applied survival analysis: regression modeling of time-to-event data. 2nd ed. Hoboken, New Jersey: John Wiley & Sons; 2011. doi: 10.1002/9780470258019.
[36] Bender R, Augustin T, Blettner M. Generating survival times to simulate cox proportional hazards models. Stat Med. 2005;24(11):1713–23. https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2059.
[37] Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. Epidemiology. 2016;27(3):368–77. doi: 10.1097/EDE.0000000000000457.
[38] VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the e-value. Ann Internal Med. 2017;167(4):268–74. https://www.acpjournals.org/doi/abs/10.7326/M16-2607.
[39] Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. JAMA Cardiol. 2017;2(11):1179–80.
[40] Tian L, Jin H, Uno H, Lu Y, Huang B, Anderson KM, et al. On the empirical choice of the time window for restricted mean survival time. Biometrics. 2020;76(4):1157–66.
[41] Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. Biometrics. 1973;29(1):185–203. http://www.jstor.org/stable/2529685.

# Appendix
# A Theoretical results in Section 2

We will prove the propositions and related lemmas in Section 2.

**Proposition A1.** *Given Assumptions* 1–2, *we have* $(T^0, T^1) \perp\!\!\!\perp A|e(X)$, *which further implies* $(Z^0, Z^1) \perp\!\!\!\perp A|e(X)$.

**Proof.** It is equivalent to show $P\{A = 1|T^1, T^0, e(X)\} = P\{A = 1|e(X)\}$.

By Theorem 2 in Rosenbaum and Rubin [4], we have $P(A = 1|e(X)) = E\{e(X)|e(X)\} = e(X)$, then it is equivalent to show $P\{A = 1|T^1, T^0, e(X)\} = e(X)$. We have

$$\begin{aligned} P\{A = 1|T^1, T^0, e(X)\} &= E\{P(A = 1|T^1, T^0, X)|T^1, T^0, e(X)\} \\ &= E\{P(A = 1|X)|T^1, T^0, e(X)\} \text{(by strongly ignorability)} \\ &= E\{e(X)|T^1, T^0, e(X)\} = e(X) = P\{A = 1|e(X)\}. \end{aligned}$$

Thus, we have $(T^0, T^1) \perp\!\!\!\perp A|e(X)$ for $0 < \text{pr}(A = 1|e(X)) < 1$. Since $Z^A = \min(T^A, \tau)$ and $\tau$ is a fixed constant, the aforementioned conditional independence also implies $(Z^0, Z^1) \perp\!\!\!\perp A|e(X)$. □

**Lemma A1.** *Given Assumptions* 1–3, $(T^1, T^0) \perp\!\!\!\perp A$ *holds marginally in the matched sample under the propensity score matching design.*

**Proof.** By Assumption 2, we have $(T^1, T^0) \perp\!\!\!\perp A|e(X)$, where $0 < P(A = 1|e(X)) < 1$. Let $M$ denotes the matching structure, and $\varepsilon_M$ denotes the set of propensity scores in the matched sample. Then, we have the following equation by matching on propensity score $e(X)$ with a constant treatment to control allocation ratio $1 : k$ ($k = 1$ for pair matching),

$$P(A = 1|e(X)) = \frac{1}{k+1}, \quad \text{for all } e(X) \in \varepsilon_M.$$

Thus, $e(X) \perp\!\!\!\perp A$ holds in the matched sample, i.e., $f_M(e(X)|A) = f_M(e(X))$. Consider the joint density of $T^1$ and $T^0$ conditional on $A$ in the matched sample, which is denoted as $f_M(T^1, T^0|A)$, we have

$$\begin{aligned} f_M(T^1, T^0|A) &= \int_{\varepsilon_M} f(T^1, T^0|A, e(X)) f_M(e(X)|A) de(X) \\ &= \int_{\varepsilon_M} f(T^1, T^0|A, e(X)) f_M(e(X)) de(X) \quad \text{[matched by constant allocation ratio]} \\ &= \int_{\varepsilon_M} f(T^1, T^0|e(X)) f_M(e(X)) de(X) \quad \text{[by Assumption 2]} \\ &= f_M(T^1, T^0). \end{aligned}$$

Since $f_M(T^1, T^0|A) = f_M(T^1, T^0)$ implies $(T^1, T^0) \perp\!\!\!\perp A$ in the matched sample, $(T^1, T^0) \perp\!\!\!\perp A$ holds marginally in the matched sample. □

**Lemma A2.** *Let* $\hat{S}_{e(X),A}(t)$ *denotes the KM survival function estimator given propensity score* $e(X)$ *and treatment indicator* $A$. *For a fixed truncation time* $\tau$,

$$\lim_{n\to\infty} \int_0^\tau E_T[\hat{S}_{e(X),A}(t) - S_{e(X),A}(t)]\mathrm{d}t = 0,$$

**Proof.** Define $\tilde{T} = \min(T, C)$ and $\pi_{e(X),A}(t) = P(\tilde{T} \geq t) \in (0, 1)$, then $[1 - \pi_{e(X),A}(t)]^n$ is a nonnegative function that increases as $t$ increases. By Lemma 3.2.1 in the study by Fleming and Harrington [29], we know:

$$\int_0^\tau E_T[\hat{S}_{e(X),A}(t) - S_{e(X),A}(t)]\mathrm{d}t \le \int_0^\tau [1 - S_{e(X),A}(t)][1 - \pi_{e(X),A}(t)]^n \mathrm{d}t$$
$$\le \int_0^\tau [1 - \pi_{e(X),A}(t)]^n \mathrm{d}t \le \tau[1 - \pi_{e(X),A}(\tau)]^n.$$

Since $\tau > 0$ is a fixed constant and $1 - \pi_{e(X),A}(\tau) \in (0, 1)$, we have

$$\lim_{n\to\infty} \int_0^\tau E_T[\hat{S}_{e(X),A}(t) - S_{e(X),A}(t)]\mathrm{d}t \le \lim_{n\to\infty} \tau[1 - \pi_{e(X),A}(\tau)]^n = 0.$$

Therefore, we have $\lim_{n\to\infty}\int_0^\tau E_T[\hat{S}_{e(X),A}(t) - S_{e(X),A}(t)]\mathrm{d}t = 0$. □

**Proposition A2.** *Given Assumptions* 1–4, *the RMST estimator based on the KM method given propensity score* $e(X)$ *and treatment group* $A$, *denoted as* $\hat{\mu}_{e(X),A}$, *is an asymptotically unbiased estimator for* $\mu_{e(X),A}$ *given* $\tau < t_{\max}$.

**Proof.** First, we will show that $\hat{S}_{e(X),A}(t)$ is asymptotically unbiased for any time $T < t_{\max}$. Let $t_i$'s be i.i.d event times ranking from small to large, and $Y_i$ is the number of people at risk at event time $t_i$. Let $d_i$ be the number of event at event time $t_i$, then we have the following definition.

$$\hat{S}_{e(X),A}(t) = \begin{cases} 1, & \text{if } t \le t_1 \quad \text{given } e(X) \text{ and } A \\ \prod_{t_i \le t} \frac{Y_i - d_i}{Y_i}, & \text{if } t_1 \le t \text{ given } e(X) \text{ and } A. \end{cases}$$

Let $\hat{\Lambda}_{e(X),A}(u) = \sum_{t_i<t} \frac{d_i}{Y_i}$ be the Nelson–Aalen estimator for the cumulative hazard function $\Lambda_{e(X),A}(u)$ given $e(X)$ and $A$. According to Theorem 3.2.3 in the study by Fleming and Harrington [29], we have the following equation if $S_{e(X),A}(t) > 0$:

$$\frac{\hat{S}_{e(X),A}(t)}{S_{e(X),A}(t)} - 1 = -\int_0^t \frac{\hat{S}_{e(X),A}(u^-)}{S_{e(X),A}(u)} d\{\hat{\Lambda}_{e(X),A}(u) - \Lambda_{e(X),A}(u)\},$$
$$E[\hat{S}_{e(X),A}(t) - S_{e(X),A}(t)] = E\left[ I_{\{T<t\}} \frac{\hat{S}_{e(X),A}(T)\{S_{e(X),A}(T) - S_{e(X),A}(t)\}}{S_{e(X),A}(T)} \right].$$

Based on Lemma 3.2.1 in the study by Fleming and Harrington [29], the bias $E[\hat{S}_{e(X),A}(t) - S_{e(X),A}(t)]$ will converge to zero as sample size $n \to \infty$. Thus, $\hat{S}_{e(X),A}(t)$ is asymptotically unbiased given $t < t_{\max}$. Similarly, $\hat{S}_{e(X),A=0}(t)$ is also asymptotically unbiased given $t < t_{\max}$.

Second, we will show that $\hat{\mu}_{e(X),A}$ is an asymptotically unbiased estimator given $\tau < t_{\max}$. Since $E_T(\hat{\mu}_{e(X),A}) = E_T\left[\int_0^\tau \hat{S}_{e(X),A}(t)\mathrm{d}t\right]$ and $\hat{S}_{e(X),A}(t)$ is a positive bounded function between 0 and 1 when $t \in [0, \tau]$, then we have

$$E_T\left[\int_0^\tau |\hat{S}_{e(X),A}(t)|\mathrm{d}t\right] = E_T\left[\int_0^\tau \hat{S}_{e(X),A}(t)\mathrm{d}t\right] \le \tau < \infty.$$

By Fubini's theorem,

$$E_T\left[\int_0^\tau \hat{S}_{e(X),A}(t)\mathrm{d}t\right] = \int_0^\tau E_T[\hat{S}_{e(X),A}(t)]\mathrm{d}t.$$

By Proposition A2 and Lemma A2, we have the following for fixed truncated time $\tau$.

$$\begin{aligned}
\lim_{n\to\infty} E_T(\hat{\mu}_{e(X),A}) - \mu_{e(X),A} &= \lim_{n\to\infty} E_T\left[\int_0^\tau \hat{S}_{e(X),A}(t)\mathrm{d}t\right] - \int_0^\tau S_{e(X),A}(t)\mathrm{d}t \\
&= \lim_{n\to\infty} \int_0^\tau E_T[\hat{S}_{e(X),A}(t)]\mathrm{d}t - \int_0^\tau S_{e(X),A}(t)\mathrm{d}t \quad \text{(by Fubini's theorem)} \\
&= \lim_{n\to\infty} \int_0^\tau E_T\{\hat{S}_{e(X),A}(t) - S_{e(X),A}(t)\}\mathrm{d}t \\
&\le \lim_{n\to\infty} \tau[1-\pi_{e(X),A}(\tau)]^n = 0.
\end{aligned}$$

Therefore, $\hat{\mu}_{e(X),A}$ is an asymptotically unbiased estimator for $\mu_{e(X),A}$ when $\tau < t_{\max}$. □

**Lemma A3.** *For a fixed truncation time point* $\tau < t_{\max}$,

$$\lim_{n\to\infty} \int_0^\tau E_T[\hat{S}^A A(t) - S^A(t)]\mathrm{d}t = 0.$$

**Proof.** Let $\tilde{T} = \min(T, C)$ and $\pi(t) = P(\tilde{T} \ge t) \in (0, 1)$, then $[1-\pi(t)]^n$ is a nonnegative function, which increases as $t$ increases. Let $\pi_A(t)$ denotes the function $\pi(t)$ for treatment indicator A. By Lemma 3.2.1 in the study by Fleming and Harrington [29], we have

$$\int_0^\tau E_T[\hat{S}^A(t) - S^A(t)]\mathrm{d}t \le \int_0^\tau [1-S^A(t)][1-\pi_A(t)]^n\mathrm{d}t \le \int_0^\tau [1-\pi_A(t)]^n\mathrm{d}t \le \tau[1-\pi_A(t)]^n.$$

Since $\tau > 0$ is a fixed constant and $1 - \pi_A(\tau) \in (0, 1)$, we have

$$\lim_{n\to\infty} \int_0^\tau E_T[\hat{S}^A(t) - S^A(t)]\mathrm{d}t \le \lim_{n\to\infty} \tau[1-\pi_A(t)]^n = 0.$$ □

**Proposition A3.** *Given Assumptions* 1–4, $\hat{\Delta}_{\mathrm{ATT}} = \int_0^\tau [\hat{S}^1(t) - \hat{S}^0(t)]\mathrm{d}t$ *is asymptotically unbiased.*

**Proof.** Since the estimated survival function $\hat{S}(t) \in (0, 1)$ and $\left|\int_0^\tau \hat{S}(t)\mathrm{d}t\right| \in (0, \tau)$, we also satisfy the following conditions to use Fubini's theorem:

(1) $E_{e(X)}\left\{E_T\left|\int_0^\tau \hat{S}_{e(X),A}(t)\mathrm{d}t\right|\right\} \le E_{e(X)}\{E_T(\tau)\} = \tau < \infty$

(2) $E_{e(X)}\left[\int_0^\tau |\hat{S}_{e(X),A}(t)|\mathrm{d}t\right] \le \tau < \infty$

(3) $E_T\left\{\int_0^\tau \hat{S}^A(t)\mathrm{d}t\right\} \le \tau < \infty$

Thus, we can apply Fubini's theorem three times to interchange the expectation of $e(X)$ as follows:

$$
\begin{aligned}
& E_{e(X)}\left\{E_T\left[\int_0^\tau \hat{S}_{e(X),A=1}(t)\mathrm{d}t\right] - E_T\left[\int_0^\tau \hat{S}_{e(X),A=0}(t)\mathrm{d}t\right]\right\} \\
&= E_T\left\{E_{e(X)}\left[\int_0^\tau \hat{S}_{e(X),A=1}(t)\mathrm{d}t\right] - E_{e(X)}\left[\int_0^\tau \hat{S}_{e(X),A=0}(t)\mathrm{d}t\right]\right\} \\
&= E_T\left\{\int_0^\tau E_{e(X)}[\hat{S}_{e(X),A=1}(t)]\mathrm{d}t - \int_0^\tau E_{e(X)}[\hat{S}_{e(X),A=0}(t)]\mathrm{d}t\right\} \\
&= E_T\left\{\int_0^\tau \hat{S}^1(t)\mathrm{d}t - \int_0^\tau \hat{S}^0(t)\mathrm{d}t\right\} \\
&= \int_0^\tau E_T[\hat{S}^1(t)]\mathrm{d}t - \int_0^\tau E_T[\hat{S}^0(t)]\mathrm{d}t.
\end{aligned}
$$

By Lemma A3, we have

$$
\begin{aligned}
& \lim_{n\to\infty} E_{e(X)}\left\{E_T\left[\int_0^\tau \hat{S}_{e(X),A=1}(t)\mathrm{d}t\right] - E_T\left[\int_0^\tau \hat{S}_{e(X),A=0}(t)\mathrm{d}t\right]\right\} \\
&= \lim_{n\to\infty} \int_0^\tau E_T[\hat{S}^1(t)]\mathrm{d}t - \lim_{n\to\infty} \int_0^\tau E_T[\hat{S}^0(t)]\mathrm{d}t \\
&= \int_0^\tau E_T[S^1(t)]\mathrm{d}t - \int_0^\tau E_T[S^0(t)]\mathrm{d}t = \mu_1 - \mu_0.
\end{aligned}
$$

Therefore, our proposed propensity score matched RMST estimator is asymptotically unbiased when truncation time point $\tau < t_{\max}$. □

# B Theoretical results in Section 4

## B.1 Proofs of propositions about conditional effect

We define the expectations of the RMST outcome $Z = \min(T, \tau)$. The following propositions are proved **within each propensity score value** $e(X) = e(x)$.

**Proposition A4.** *For binary unmeasured confounder* $U = 0, 1$, *we have* $\mathrm{RR}_{AU|e(X)=e(x)} \geq 1$ *and* $\mathrm{MR}_{UZ|e(X)=e(x)} \geq 1$.

**Proof.** By definition, we have $\mathrm{MR}_{UZ|A=1,e(X)=e(x)} \geq 1$ and $\mathrm{MR}_{UZ|A=0,e(X)=e(x)} \geq 1$, then $\mathrm{MR}_{UZ|e(X)=e(x)} = \max(\mathrm{MR}_{UZ|A=1,e(X)=e(x)}, \mathrm{MR}_{UZ|A=0,e(X)=e(x)}) \geq 1$. Assume $\mathrm{RR}_{AU|e(x)} = \max_{k=0,1}\mathrm{RR}_{AU,k|e(x)} < 1$, then it implies that

$$
\begin{aligned}
P(U = 0|A = 1, e(X) = e(x)) < P(U = 0|A = 0, e(X) = e(x)), \\
P(U = 1|A = 1, e(X) = e(x)) < P(U = 1|A = 0, e(X) = e(x)).
\end{aligned}
$$

This further implies that $1 = P(U = 0|A = 1, e(X) = e(x)) + P(U = 1|A = 1, e(X) = e(x)) < P(U = 0|A = 0, e(X) = e(x)) + P(U = 1|A = 0, e(X) = e(x)) = 1$, which is not true. Thus, we have proved by contradiction that $\mathrm{RR}_{AU|e(X)=e(x)} \geq 1$. □

**Proposition A5.**

$$\mathrm{CMR}_{AZ^+} = \frac{\mathrm{MR}_{AZ}}{\mathrm{MR}^{\mathrm{true}}_{AZ^+}} \le BF_U, \quad \mathrm{CMR}_{AZ^-} = \frac{\mathrm{MR}_{AZ}}{\mathrm{MR}^{\mathrm{true}}_{AZ^-}} \le BF_U, \quad \mathrm{CMR}_{AZ} = \frac{\mathrm{MR}_{AZ}}{\mathrm{MR}^{\mathrm{true}}_{AZ}} \le BF_U.$$

**Proof.** First, let $f = P(A = 1)$, then we have

$$\mathrm{MR}^{\mathrm{true}}_{AZ} = \frac{\int r_1(u)F(\mathrm{d}u)}{\int r_0(u)F(\mathrm{d}u)} = \frac{f\int r_1(u)F_1(\mathrm{d}u) + (1-f)\int r_1(u)F_0(\mathrm{d}u)}{f\int r_0(u)F_1(\mathrm{d}u) + (1-f)\int r_0(u)F_0(\mathrm{d}u)} \tag{A1}$$

$$= \frac{f\int r_0(u)F_1(\mathrm{d}u)}{f\int r_0(u)F_1(\mathrm{d}u) + (1-f)\int r_0(u)F_0(\mathrm{d}u)} \times \frac{\int r_1(u)F_1(\mathrm{d}u)}{\int r_0(u)F_1(\mathrm{d}u)} \tag{A2}$$

$$+ \frac{(1-f)\int r_0(u)F_0(\mathrm{d}u)}{f\int r_0(u)F_1(\mathrm{d}u) + (1-f)\int r_0(u)F_0(\mathrm{d}u)} \times \frac{\int r_1(u)F_0(\mathrm{d}u)}{\int r_0(u)F_0(\mathrm{d}u)}. \tag{A3}$$

Let $w = \frac{f\int r_0(u)F_1(\mathrm{d}u)}{f\int r_0(u)F_1(\mathrm{d}u) + (1-f)\int r_0(u)F_0(\mathrm{d}u)} \in [0,1]$, then we have

$$\mathrm{MR}^{\mathrm{true}}_{AZ} = w\mathrm{MR}^{\mathrm{true}}_{AZ^+} + (1-w)\mathrm{MR}^{\mathrm{true}}_{AZ^-}; \quad \frac{1}{\mathrm{CMR}_{AZ}} = \frac{w}{\mathrm{CMR}_{AZ^+}} + \frac{1-w}{\mathrm{CMR}_{AZ^-}}.$$

Second, we have

$$\mathrm{CMR}_{AZ^+} = \frac{\mathrm{MR}^{\mathrm{obs}}_{AZ}}{\mathrm{MR}^{\mathrm{true}}_{AZ^+}} = \frac{\int r_1(u)F(\mathrm{d}u)}{\int r_0(u)F(\mathrm{d}u)} \Big/ \frac{\int r_1(u)F_1(\mathrm{d}u)}{\int r_0(u)F_1(\mathrm{d}u)} = \frac{w_1 \max_u r_0(u) + (1-w_1)\min_u r_0(u)}{w_0 \max_u r_0(u) + (1-w_0)\min_u r_0(u)},$$

where $w_1 = \frac{\int [r_0(u) - \min_u r_0(u)]F_1(\mathrm{d}u)}{\max_u r_0(u) - \min_u r_0(u)}$ and $w_0 = \frac{\int [r_0(u) - \min_u r_0(u)]F_0(\mathrm{d}u)}{\max_u r_0(u) - \min_u r_0(u)}$.

Define $\Gamma = \frac{w_1}{w_0}$, then

$$\Gamma = \frac{w_1}{w_0} = \frac{\int [r_0(u) - \min_u r_0(u)]F_1(\mathrm{d}u)}{\int [r_0(u) - \min_u r_0(u)]F_0(\mathrm{d}u)} = \frac{\int [r_0(u) - \min_u r_0(u)]\mathrm{RR}_{AU}(u)F_0(\mathrm{d}u)}{\int [r_0(u) - \min_u r_0(u)]F_0(\mathrm{d}u)} \tag{A4}$$

$$\le \frac{\max_x \mathrm{RR}_{AU}(u)\int [r_0(u) - \min_u r_0(u)]F_0(\mathrm{d}u)}{\int [r_0(u) - \min_u r_0(u)]F_0(\mathrm{d}u)} = \mathrm{RR}_{AU}. \tag{A5}$$

Write $w_0 = \frac{w_1}{\Gamma}$, then

$$\mathrm{CMR}^+_{AZ} = \frac{[\max_u r_0(u) - \min_u r_0(u)]w_1 + \min_u r_0(u)}{[\max_x r_0(u) - \min_u r_0(u)]w1\Gamma + \min_u r_0(u)}.$$

If $\Gamma > 1$, $\mathrm{CMR}^+_{AZ}$ is increasing in $w_1$ according to Lemma A.1 in the Appendix of Ding and VanderWeele [37], then the maximum attains at $w_1 = 1$, and we have

$$\mathrm{CMR}^+_{AZ} \le \frac{\Gamma \times \mathrm{MR}_{UZ|A=0}}{\Gamma + \mathrm{MR}_{UZ|A=0} - 1} \le \frac{\mathrm{RR}_{AU} \times \mathrm{MR}_{UZ|A=0}}{\mathrm{RR}_{AU} + \mathrm{MR}_{UZ|A=0} - 1}.$$

If $\Gamma \le 1$, $\mathrm{CMR}^+_{AZ}$ is nonincreasing in $w_1$ according to Lemma A.1 in the Appendix of Ding and VanderWeele [37], then the maximum attains at $w_1 = 0$, and we have

$$\mathrm{CMR}^+_{AZ} \le 1 \le \frac{\mathrm{RR}_{AU} \times \mathrm{MR}_{UZ|A=0}}{\mathrm{RR}_{AU} + \mathrm{MR}_{UZ|A=0} - 1}.$$

Similarly, by $\frac{1}{\mathrm{CMR}_{AZ}} = \frac{w}{\mathrm{CMR}_{AZ^+}} + \frac{1-w}{\mathrm{CMR}_{AZ^-}}$, we have

$$\frac{1}{\mathrm{CMR}_{AZ}} \ge \frac{1}{BF_U}, \quad \mathrm{CMR}_{AZ} \le BF_U.$$

□

To study the average causal effect of the exposure on the difference scale, we need the following definitions:

- Define $m_0 = E(Z|A = 0)$ and $m_1 = E(Z|A = 1)$, then the observed mean difference of exposure on the outcome is $m_1 - m_0$.
- The average causal effect of the exposure on the outcome for exposed is

$$\mathrm{ACE}_{AZ^+}^{\mathrm{true}} = \int E(Z|A = 1, U = u)F_1(\mathrm{d}u) - \int E(Z|A = 0, U = u)F_1(\mathrm{d}u) = m_1 - \int r_0(u)F_1(\mathrm{d}u).$$

- The average causal effect of the exposure on the outcome for unexposed is expressed as follows:

$$\mathrm{ACE}_{AZ^-}^{\mathrm{true}} = \int E(Z|A = 1, U = u)F_0(\mathrm{d}u) - \int E(Z|A = 0, U = u)F_0(\mathrm{d}u) = \int r_1(u)F_0(\mathrm{d}u) - m_0.$$

- The average causal effect of the exposure on the outcome for whole population is

$$\mathrm{ACE}_{AZ}^{\mathrm{true}} = \int E(Z|A = 1, U = u)F(\mathrm{d}u) - \int E(Z|A = 0, U = u)F(\mathrm{d}u) = f\,\mathrm{ACE}_{AZ^+}^{\mathrm{true}} + (1 - f)\mathrm{ACE}_{AZ^-}^{\mathrm{true}}.$$

**Proposition A6.** *For nonnegative outcomes and* $\mathrm{ACE}_{AZ}^{\mathrm{obs}} \geq 0$, *the lower bounds for the average causal effects are expressed as follows:*

$$\mathrm{ACE}_{AZ^+}^{\mathrm{true}} \geq m_1 - m_0 \times BF_U; \quad \mathrm{ACE}_{AZ^-}^{\mathrm{true}} \geq m_1 / BF_U - m_0;$$
$$\mathrm{ACE}_{AZ}^{\mathrm{true}} \geq (m_1 - m_0 \times BF_U)[f + (1 - f)/BF_U] = \left(\frac{m_1}{BF_U} - m_0\right)[f \times BF_U + (1 - f)].$$

**Proof.** From the data, we can identify

$$m_1 = \int E(Z|A = 1, U = u)F_1(\mathrm{d}u) = \int r_1(u)F_1(\mathrm{d}u) = E(Z|A = 1);$$
$$m_0 = \int E(Z|A = 0, U = u)F_0(\mathrm{d}u) = \int r_0(u)F_0(\mathrm{d}u) = E(Z|A = 0).$$

The counterfactual probabilities are not identifiable:

$$E(Z(1) = 1|A = 0) = \int E(Z = 1|A = 1, U = u)F_0(\mathrm{d}u) = \int r_1(u)F_0(\mathrm{d}u);$$
$$E(Z(0) = 1|A = 1) = \int E(Z = 1|A = 0, U = u)F_1(\mathrm{d}u) = \int r_0(u)F_1(\mathrm{d}u).$$

First, by Proposition A5, we have

$$\frac{m_1}{E(Z(1) = 1|A = 0)} = \frac{\int r_1(u)F_1(\mathrm{d}u)}{\int r_1(u)F_0(\mathrm{d}u)} = \frac{\int r_1(u)F_1(\mathrm{d}u)}{\int r_0(u)F_0(\mathrm{d}u)} \Big/ \frac{\int r_1(u)F_0(\mathrm{d}u)}{\int r_0(u)F_0(\mathrm{d}u)} = \frac{\mathrm{MR}_{AZ}}{\mathrm{MR}_{AZ^-}^{\mathrm{true}}} = \mathrm{CMR}_{AZ^-} \leq BF_U.$$

Thus, we have $E(Z(1 = 1)|A = 0) \geq \frac{m_1}{BF_U}$.

Second, by Proposition A5 again, we have

$$\frac{E(Z(0) = 1|A = 1)}{m_0} = \frac{\int r_0(u)F_1(\mathrm{d}u)}{\int r_0(u)F_0(\mathrm{d}u)} = \mathrm{CMR}_{AZ^+} \leq BF_U.$$

Thus, we have $E(Z(0) = 1|A = 1) \leq m_0 BF_U$.

By definition of ACE and the inequalities derived earlier, we have

$$\mathrm{ACE}_{AZ^+}^{\mathrm{true}} = m_1 - \int r_0(u)F_1(\mathrm{d}u) \geq m_1 - m_0 \times BF_U;$$
$$\mathrm{ACE}_{AZ^-}^{\mathrm{true}} = \int r_1(u)F_0(\mathrm{d}u) - m_0 \geq m_1 / BF_U - m_0;$$

$$\begin{aligned}\mathrm{ACE}_{AZ}^{\mathrm{true}} &= f \cdot \mathrm{ACE}_{AZ^+}^{\mathrm{true}} + (1-f)\mathrm{ACE}_{AZ^-}^{\mathrm{true}} \\ &\geq f(m_1 - m_0 BF_U) + (1-f)\left(\frac{m_1}{BF_U} - m_0\right) \\ &= (m_1 - m_0 \times BF_U)[f + (1-f)/BF_U] \\ &= \left(\frac{m_1}{BF_U} - m_0\right)[f \times BF_U + (1-f)].\end{aligned}$$

□

**Proposition A7.** *For nonnegative outcomes with* $\mathrm{ACE}_{AZ}^{\mathrm{obs}} < 0$, *we have*

$$\mathrm{ACE}_{AZ^+}^{\mathrm{true}} \leq m_1 BF_U - m_0; \quad \mathrm{ACE}_{AZ^-}^{\mathrm{true}} \leq m_1 - \frac{m_0}{BF_U};$$

$$\mathrm{ACE}_{AZ}^{\mathrm{true}} \leq (m_1 BF_U - m_0)\left(f + \frac{1-f}{BF_U}\right) = \left(m_1 - \frac{m_0}{BF_U}\right)(fBF_U + 1 - f).$$

**Proof.** Define $\bar{A} = 1 - A$. By applying Proposition A6 we have

$$\begin{aligned}&\mathrm{ACE}_{\bar{A}Z^+}^{\mathrm{true}} \geq E(Z|\bar{A}=1) - E(Z|\bar{A}=0) \times BF_U; \\ &\mathrm{ACE}_{\bar{A}Z^-}^{\mathrm{true}} \geq E(Z|\bar{A}=1)/BF_U - E(Z|\bar{A}=0); \\ &\mathrm{ACE}_{\bar{A}Z}^{\mathrm{true}} \geq (E(Z|\bar{A}=1) - E(Z|\bar{A}=0) \times BF_U)[f + (1-f)/BF_U] \\ &\qquad = \left(\frac{E(Z|\bar{A}=1)}{BF_U} - E(Z|\bar{A}=0)\right)[f \times BF_U + (1-f)].\end{aligned}$$

Because $\mathrm{ACE}_{\bar{A}Z^+}^{\mathrm{true}} = -\mathrm{ACE}_{AZ^+}^{\mathrm{true}}$, $\mathrm{ACE}_{\bar{A}Z^-}^{\mathrm{true}} = -\mathrm{ACE}_{AZ^-}^{\mathrm{true}}$, and $\mathrm{ACE}_{\bar{A}Z}^{\mathrm{true}} = -\mathrm{ACE}_{AZ}^{\mathrm{true}}$, and we also have $E(Z|\bar{A}=0) = E(Z|A=1) = m_1$ and $E(Z|\bar{A}=1) = E(Z|A=0) = m_0$. Then we have

$$\mathrm{ACE}_{AZ^+}^{\mathrm{true}} \leq m_1 BF_U - m_0; \quad \mathrm{ACE}_{AZ^-}^{\mathrm{true}} \leq m_1 - \frac{m_0}{BF_U};$$

$$\mathrm{ACE}_{AZ}^{\mathrm{true}} \leq (m_1 BF_U - m_0)\left(f + \frac{1-f}{BF_U}\right) = \left(m_1 - \frac{m_0}{BF_U}\right)(fBF_U + 1 - f).$$

□

## B.2 Proofs of propositions about the marginal effect

To make the bounding factor hold for all propensity score values, we consider the maximum value of $BF_U$ across all values of propensity score $e(X)$, which is defined as $BF_U^* = \max_{e(x)}(BF_{U|e(X)=e(x)})$.

**Proposition A8.** *For nonnegative outcomes and* $\mathrm{ACE}_{AZ}^{\mathrm{obs}} \geq 0$, *we have*

$$\mathrm{ACE}_{AZ^+}^{\mathrm{true}} \geq m_1 - m_0 \times BF_U^*; \quad \mathrm{ACE}_{AZ^-}^{\mathrm{true}} \geq m_1/BF_U^* - m_0;$$

$$\mathrm{ACE}_{AZ}^{\mathrm{true}} \geq (m_1 - m_0 \times BF_U^*)[f + (1-f)/BF_U^*] = \left(\frac{m_1}{BF_U^*} - m_0\right)[f \times BF_U^* + (1-f)].$$

*For nonnegative outcomes and* $\mathrm{ACE}_{AZ}^{\mathrm{obs}} < 0$, *we have*

$$\mathrm{ACE}_{AZ^+}^{\mathrm{true}} \leq m_1 BF_U^* - m_0;$$

$$\mathrm{ACE}_{AZ^-}^{\mathrm{true}} \leq m_1 - \frac{m_0}{BF_U^*};$$

$$\mathrm{ACE}_{AZ}^{\mathrm{true}} \leq (m_1 BF_U^* - m_0)\left(f + \frac{1-f}{BF_U^*}\right) = \left(m_1 - \frac{m_0}{BF_U^*}\right)(fBF_U^* + 1 - f).$$

**Proof.** We will start from showing the results for nonnegative outcomes and $\text{ACE}_{AZ}^{\text{obs}} \geq 0$. First, we have

$$\frac{m_1}{E(Z(1)=1|A=0)} = \frac{\int r_1(u)F_1(\mathrm{d}u)}{\int r_1(u)F_0(\mathrm{d}u)} = \frac{\int r_1(u)F_1(\mathrm{d}u)}{\int r_0(u)F_0(\mathrm{d}u)} \Big/ \frac{\int r_1(u)F_0(\mathrm{d}u)}{\int r_0(u)F_0(\mathrm{d}u)}$$
$$= \frac{\text{MR}_{AZ}}{\text{MR}_{AZ^-}^{\text{true}}} = \text{CMR}_{AZ^-} \leq BF_U \leq BF_U^*.$$

Thus, we have $E(Z(1)=1|A=0) \geq \frac{m_1}{BF_U} \geq \frac{m_1}{BF_U^*}$.

Second, we know that $\frac{E(Z(0)=1 \mid A=1)}{m_0} = \frac{\int r_0(u)F_1(\mathrm{d}u)}{\int r_0(u)F_0(\mathrm{d}u)} = \text{CMR}_{AZ^+} \leq BF_U \leq BF_U^*$, then we have $E(Z(0)=1|A=1) \leq m_0 BF_U \leq m_0 BF_U^*$.

By definition of ACE and the inequalities derived earlier, we have

$$\begin{aligned}
\text{ACE}_{AZ^+}^{\text{true}} &= m_1 - \int r_0(u)F_1(\mathrm{d}u) \geq m_1 - m_0 \times BF_U^*;\\
\text{ACE}_{AZ^-}^{\text{true}} &= \int r_1(u)F_0(\mathrm{d}u) - m_0 \geq m_1/BF_U^* - m_0;\\
\text{ACE}_{AZ}^{\text{true}} &= f \cdot \text{ACE}_{AZ^+}^{\text{true}} + (1-f)\text{ACE}_{AZ^-}^{\text{true}}\\
&\geq f(m_1 - m_0 BF_U^*) + (1-f)\left(\frac{m_1}{BF_U^*} - m_0\right)\\
&= (m_1 - m_0 \times BF_U^*)[f + (1-f)/BF_U^*]\\
&= \left(\frac{m_1}{B^*F_U} - m_0\right)[f \times BF_U^* + (1-f)].
\end{aligned}$$

Similarly, we can prove the inequalities hold for nonnegative outcomes and $\text{ACE}_{AZ}^{\text{obs}} < 0$. □

**Proposition A9.** *In the matched sample, we have the following inequality for nonnegative outcomes and* $\text{ACE}_{AZ}^{\text{obs}} \geq 0$:

$$\text{ACE}_{AZ}^{\text{true}} \geq \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right)E(Z|A=1) - \frac{1}{2}(1 + BF_U^*)E(Z|A=0).$$

*In the matched sample, we have the following inequality for nonnegative outcomes and* $\text{ACE}_{AZ}^{\text{obs}} < 0$:

$$\text{ACE}_{AZ}^{\text{true}} \leq \frac{1}{2}(1 + BF_U^*)E(Z|A=1) - \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right)E(Z|A=0).$$

**Proof.** In the matched sample, we have $f = P(A=1|e(X)=e(x)) = 0.5$. For nonnegative outcomes and $\text{ACE}_{AZ}^{\text{obs}} \geq 0$, we have LHS $= \text{ACE}_{AZ}^{\text{true}} = \sum_{e(x)} \text{ACE}_{AZ|e(X)=e(x)}^{\text{true}} P(e(X)=e(x))$ and

$$\begin{aligned}
\text{RHS} &= \sum_{e(x)} (m_1 - m_0 BF_U^*)\left(f + \frac{1-f}{BF_U^*}\right)P(e(X)=e(x))\\
&= \left(\frac{1}{2} - \frac{1}{2}BF_U^*\right)\sum_{e(x)} m_1 P(e(X)=e(x)) + \frac{1}{BF_U^*}\sum_{e(x)} m_1 P(e(X)=e(x))\\
&\quad + \left(\frac{1}{2} - \frac{1}{2}BF_U^*\right)\sum_{e(x)} M_0 P(e(X)=e(x)) - \sum_{e(x)} m_0 P(e(X)=e(x))\\
&= \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right)E(Z|A=1) - \frac{1}{2}(1 + BF_U^*)E(Z|A=0).
\end{aligned}$$

Thus, we have $\text{ACE}_{AZ}^{\text{true}} \geq \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right)E(Z|A=1) - \frac{1}{2}(1 + BF_U^*)E(Z|A=0)$.

For $\mathrm{ACE}_{AZ}^{\mathrm{obs}} < 0$, we have LHS $= \mathrm{ACE}_{AZ}^{\mathrm{true}} = \sum_{e(x)} \mathrm{ACE}_{AZ|e(X)=e(x)}^{\mathrm{true}} P(e(X) = e(x))$ and

$$\begin{aligned}
\mathrm{RHS} &= \sum_{e(x)} (m_1 BF_U^* - m_0)\left(f + \frac{1-f}{BF_U^*}\right) P(e(X) = e(x)) \\
&= \sum_{e(x)} (m_1 BF_U^* - m_0)\left(\frac{1}{2} + \frac{1}{2BF_U^*}\right) P(e(X) = e(x)) \\
&= \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right)[BF_U^* \sum_{e(x)} m_1 P(e(X) = e(x)) - \sum_{e(x)} m_0 P(e(X) = e(x))] \\
&= \frac{1}{2}(1 + BF_U^*) E(Z|A = 1) - \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right) E(Z|A = 0).
\end{aligned}$$

Thus, we have $\mathrm{ACE}_{AZ}^{\mathrm{true}} \le \frac{1}{2}(1 + BF_U^*) E(Z|A = 1) - \frac{1}{2}\left(1 + \frac{1}{BF_U^*}\right) E(Z|A = 0)$. □

# C Estimation of $G_{ij}(u, v)$ in Section 2

To compute the variance of RMSTs, one difficulty is to estimate the function $G_{ij}(u, v)$ based on data. Follow the notations in the study by Murray and Cole [34], we need to transform the function $G_{ij}(u, v)$ into the counting process notation system. Suppose we have $n$ matched pairs, then let $i, j$ denote the groups and $k = 1, \ldots, n$ denotes the $k$th pair. Let $U_{ik}$ be the censoring random variable corresponding to survival time $T_{ik}$, and the censored survival time is $X_{ik} = \min(T_{ik}, U_{ik})$ with censoring status $\Delta_{ik} = I(T_{ik} < U_{ik})$. Then we have the following definitions:

(1) $Y_i(u) = \sum_{k=1}^n I(x_{ik} \ge u)$ and $Y_j(v) = \sum_{k=1}^n I(x_{jk} \ge v)$;
(2) $Y_{ij}(u, v) = \sum_{k=1}^n I(x_{ik} \ge u, x_{jk} \ge v)$;
(3) $\mathrm{d}Ni(u) = \sum_{k=1}^n I(u \le x_{ik} < u + \Delta u, \Delta_{ik} = 1)$, where $\Delta u \to 0$;
(4) $\mathrm{d}Nj(v) = \sum_{k=1}^n I(v \le x_{ik} < v + \Delta v, \Delta_{ik} = 1)$, where $\Delta v \to 0$;
(5) $\mathrm{d}N_{ij}(u, v) = \sum_{k=1}^n I(u \le x_{ik} < u + \Delta u, v \le x_{jk} < v + \Delta v, \Delta_{ik} = 1, \Delta_{jk} = 1)$, where $\Delta u \to 0$ and $\Delta v \to 0$;
(6) $\mathrm{d}N_{ij}(u|v) = \sum_{k=1}^n I(u \le x_{ik} < u + \Delta u, x_{jk} \ge v, \Delta_{ik} = 1)$, where $\Delta u \to 0$;
(7) $\mathrm{d}N_{j|i}(v|u) = \sum_{k=1}^n I(v \le x_{jk} < v + \Delta v, x_{ik} \ge u, \Delta_{jk} = 1)$, where $\Delta v \to 0$;
(8) The $\hat{G}_{ij}(u, v)$ could be estimated by the following formula, and we set $\Delta u = 0$ and $\Delta v = 0$ in real computation. The corresponding R code could be found in our supplementary materials.

$$\hat{G}_{ij}(u, v) = n \frac{Y_{ij}(u, v)}{Y_i(u) Y_j(v)} \left[ \frac{\mathrm{d}N_{ij}(u, v)}{Y_{ij}(u, v)} - \frac{\mathrm{d}N_{ij}(u|v)\mathrm{d}N_j(v)}{Y_{ij}(u, v) Y_j(v)} - \frac{\mathrm{d}N_{j|i}(v|u)\mathrm{d}N_i(u)}{Y_{ij}(u, v) Y_i(u)} + \frac{\mathrm{d}N_i(u)\mathrm{d}N_j(v)}{Y_i(u) Y_j(v)} \right].$$