Statistics in Medicine

WILEY

Statistics
in Medicine

RESEARCH ARTICLE

Adjusted Nelson–Aalen Estimators by Inverse Treatment Probability Weighting With an Estimated Propensity Score

Yuhao Deng[1] | Rui Wang[2]

[1]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA | [2]Department of Biostatistics, School of Public Health, University of Washington, Seattle, WA, USA

Correspondence: Yuhao Deng (yuhaoden@umich.edu)

Received: 9 October 2024 | **Revised:** 13 March 2025 | **Accepted:** 24 March 2025

Keywords: causal inference | competing risks | influence function | inverse probability weighting | survival analysis

ABSTRACT

Inverse probability of treatment weighting (IPW) has been well applied in causal inference to estimate population-level estimands from observational studies. For time-to-event outcomes, the failure time distribution can be estimated by estimating the cumulative hazard in the presence of random right censoring. IPW can be performed by weighting the event counting process and at-risk process by the inverse treatment probability, resulting in an adjusted Nelson–Aalen estimator for the population-level counterfactual cumulative incidence function. We consider the adjusted Nelson–Aalen estimator with an estimated propensity score in the competing risks setting. When the estimated propensity score is regular and asymptotically linear, we derive the influence functions for the counterfactual cumulative hazard and cumulative incidence. Then we establish the asymptotic properties for the estimators. We show that the uncertainty in the estimated propensity score contributes to an additional variation in the estimators. However, through simulation and real-data application, we find that such an additional variation is usually small.

1 | Introduction

For time-to-event data analysis, a key problem is that some failure times are unobserved due to right censoring. In the presence of competing risks, the cumulative incidence function (CIF) of an event is usually adopted as the estimand, which is always well defined no matter whether this event would eventually happen [1]. We may have two strategies to estimate the population-level CIF: the first is to directly estimate the incidence by weighting the observed event counting process by the inverse of the uncensored probability, and the second is to estimate the cause-specific hazards and then transform the hazards to cumulative incidence. To minimize the models imposed, the analysis for time-to-event data usually proceeds on the scale of hazards by assuming random censoring. Therefore, identifying the cause-specific hazards is the central aim in the presence of competing risks [2].

Indeed, it is challenging to specify an appropriate hazard model when there is no information on how covariates influence the hazards [3]. Modeling the post-treatment time-varying covariates is both statistically and computationally challenging, which limits the practical use [4]. Although model-based methods (e.g., Cox regression) have been well applied to estimate the coefficients of risk factors, it is not straightforward to infer the population-level cumulative incidence where the distribution of baseline covariates should be averaged out. Nonparametric approaches are sometimes preferred since there is no model restriction on the hazards, especially in randomized trials. Nonparametric estimation proceeds based on the counting processes of events, such as the Kaplan–Meier estimator using product limits and Nelson–Aalen estimator transforming the cumulative hazard [5–7]. At a tolerable cost of efficiency loss, the nonparametric approaches give consistent estimators for the cumulative incidence (or survival function) in finite sample.

A scientific question is to estimate the population-level estimand (cumulative incidence function) and average treatment effect from observational studies. In observational studies, the covariates may not be identically distributed between treatment groups. Inverse probability weighting (IPW) has been well applied in causal inference to deal with imbalanced covariates [8, 9]. An unbiased estimator for the population-level estimand is obtained by weighting the individuals in a trial using the inverse of the treatment probability. The propensity score summarizes the information of baseline covariates into a one-dimensional statistic. Conditioning on the propensity score yields balanced covariate distributions between treated and control groups. The idea of IPW can be extended to time-to-event outcomes with random right censoring. By weighting the counting processes with the inverse treatment probability, the adjusted Nelson–Aalen estimator (or asymptotically equivalently, weighted Kaplan–Meier estimator when there is only one terminal event) and weighted Aalen–Johansen estimator can be constructed [10–15]. The variance of the resulting estimator can be explicitly derived using the martingale theory if the treatment propensity score is known.

However, two problems are left unstudied in using IPW for time-to-event data. First, the nonparametric Nelson–Aalen and Aalen–Johansen estimators assume homogeneity of the hazard functions, in that all individuals in a single treatment group share the same risk of failure events [6, 7, 16]. No covariates are incorporated in the models. Second, existing work did not consider the uncertainty of the estimated propensity score [12, 17]. It is well known that using the estimated propensity score in the IPW estimator may lead to a slightly different standard deviation compared to using the true propensity score [18]. However, it is unknown how the uncertainty of the estimated propensity score will affect the inference regarding the adjusted Nelson–Aalen and Aalen–Johansen estimators.

In this article, we formalize the adjusted Nelson–Aalen estimator by inverse treatment probability weighting in the competing risks setting. We focus on the Nelson–Aalen-type estimator rather than the Aalen–Johansen-type estimator to avoid product limits in the estimators. We show the identifiability of the counterfactual cumulative incidence function for a competing event. If the propensity score is known, we construct a martingale from the counting processes. We derive the unbiasedness and the finite-sample variance of the estimated counterfactual cause-specific hazard function. If the propensity score is unknown, but the estimated propensity score is regular and asymptotically linear, we derive the influence function of the estimated counterfactual cause-specific hazard. The asymptotic property of the estimated counterfactual cumulative incidence function is then established using the empirical process theory. We find that omitting the uncertainty of the estimated propensity score may lead to a biased variance estimator. However, simulation studies show that the empirical bias is usually tiny.

The remainder of this article is organized as follows: In Section 2, we list the assumptions for identifying the counterfactual cumulative hazards and incidence functions and then give the adjusted Nelson–Aalen estimator by inverse probability of treatment weighting. In Section 3, we establish the asymptotic properties for the estimator when the propensity score is known and when the propensity score is estimated, respectively. In Section 4, we conduct simulation studies to assess the influence of using an estimated propensity score in the adjusted Nelson–Aalen estimator. We also perform a sensitivity analysis using a misspecified propensity score in the Supporting Information. In Section 5, we apply the proposed estimator to study the effect of transplant modalities on relapse and nonrelapse mortality. Finally, we end this article by pointing out some extensions in Section 6.

2 | Estimation

Let X be the baseline covariates. Let A be the treatment indicator, where $A = 1$ stands for the active treatment and $A = 0$ stands for the control (placebo). If there is only one terminal event, we let $\tilde{T}^a$ be the potential time to the event under the treatment condition $A = a$. If the event does not happen at the end of study t^*, we can denote $\tilde{T}^a > t^*$ since we are not interested in the occurrence of events after t^*. With competing events, the time to an event may not be well defined. It is more convenient to introduce the event counting processes. Under the stable unit treatment value assumption (SUTVA), let $\tilde{T}^a$ be the potential time to the first event under the treatment condition $A = a$ and let $\tilde{\Delta}^a \in \{1, \ldots, J\}$ be the potential event indicator. We denote $\tilde{T}^a_j = \tilde{T}^a$ if $\tilde{\Delta}^a = j$ and $\tilde{T}^a_j = \infty$ otherwise. Let $\tilde{N}^a_j(t) = I\{\tilde{T}^a_j \le t\}$ be the potential event counting process for the event j and $\tilde{Y}^a(t) = I\{\tilde{T}^a \ge t\}$ be the potential at-risk process. All competing events share the same at-risk process.

Subject to right censoring, we cannot fully observe the event counting processes. Let $\tilde{C}^a$ be the potential censoring time under the treatment condition $A = a$; then the potential follow-up time is $T^a = \tilde{T}^a \wedge \tilde{C}^a$. Let $\Delta^a = \tilde{\Delta}^a I\{\tilde{C}^a \ge \tilde{T}^a\}$ be the event indicator so that $\Delta^a = 0$ if censoring happens first and $\Delta^a = \tilde{\Delta}^a$ if a terminal event happens first. The event counting process for the event j with censoring is $N^a_j(t) = I\{\tilde{T}^a_j \le t, \tilde{C}^a \ge t\} = I\{T^a \le t, \Delta^a = j\}$ and the at-risk process is $Y^a(t) = I\{\tilde{T}^a \ge t, \tilde{C}^a \ge t\} = I\{T^a \ge t\}$. We adopt the ignorability assumption, which is commonly used in causal inference literature. Ignorability means that the treatment assignment is independent of the potential event time and censoring time given the baseline covariates.

Assumption 1 (Ignorability). $A \perp\!\!\!\perp (\tilde{T}^a, \tilde{\Delta}^a, \tilde{C}^a)|X$, for $a = 0, 1$.

As for censoring, we assume that the censoring is completely noninformative. The potential censoring time should be (unconditionally) independent of the potential event time.

Assumption 2 (Completely random censoring). $\tilde{C}^a \perp\!\!\!\perp (\tilde{T}^a, \tilde{\Delta}^a)$, for $a = 0, 1$.

Completely random censoring is essential for nonparametric estimation. Assumption 2 holds if the potential censoring time is independent of $(\tilde{T}^a, \tilde{\Delta}^a, X)$. In a more general case, the hazards of $\tilde{C}^a$ and $\tilde{T}^a_j$ may depend on X. Suppose that $X = (X_1, X_2)$, with $X_1 \perp\!\!\!\perp X_2$. If the hazard of $\tilde{C}^a$ is determined by X_1 but not X_2, while the hazard of $\tilde{T}^a_j$ is determined by X_2 but not X_1, then Assumption 2 holds. The backdoor paths between $\tilde{C}^a$ and $\tilde{T}^a_j$ should be blocked on the causal graph.

Define the propensity score $e(a;x) = P(A = a|X = x)$. We need positivity for the propensity score and censoring distribution.

Assumption 3 (Positivity). $0 < c < e(a;X) < 1 - c < 1$ and $P(Y^a(t^*) = 1|A = a) > c$ for $a = 0, 1$ and a constant $c > 0$.

The first part of positivity says that the propensity score is bounded away from 0 and 1. The second part of positivity says that there are still individuals at risk at the end of study, which guarantees the hazard function can be well defined in $[0, t^*]$. The probability that censoring has not happened at time t^* should be positive, so there are available data to estimate the counterfactual hazards.

In the realized trial, let T be the observed event time and Δ be the observed event indicator. We assume consistency to link the potential values with observed values.

Assumption 4 (Consistency). $T = T^A, \Delta = \Delta^A$.

The observed event counting process $N_j(t) = I\{T \le t, \Delta = j\}$ for the event j and the observed at-risk process $Y(t) = I\{T \wedge C \ge t\}$. Suppose we have a sample including n independent individuals, randomly drawn from a super-population. When necessary, we use the subscript $i = 1, \ldots, n$ to represent the individual index. The observed data can be written as either $\{(A_i, X_i, T_i, \Delta_i) : i = 1, \ldots, n\}$ or $\{(A_i, X_i, N_{ij}(s), Y_i(s), 0 \le s \le t^*) : i = 1, \ldots, n\}$.

The counterfactual hazard of event j describes the instantaneous risk of event j under the treatment condition a,

$$\begin{aligned} d\Lambda_j^a(t) &= P(t \le T_j^a < t + dt | T^a \ge t) \\ &= P(d\tilde{N}_j^a(t) = 1 | Y^a(t) = 1),\ 0 \le t < t^* \end{aligned}$$

with $dt \to 0$. By transforming the hazards, the counterfactual cumulative incidence function (CIF) of event j under the treatment condition a is given by ($a = 0, 1$, $j = 1, \ldots, J$)

$$\begin{aligned} F_j^a(t) &= P(\tilde{T}^a \le t, \tilde{\Delta}^a = j) \\ &= \int_0^t \exp\left\{ -\sum_{k=1}^{J} \Lambda_k^a(s) \right\} d\Lambda_j^a(s),\ 0 \le t < t^* \end{aligned} \tag{1}$$

We see that the hazards play the key role for time-to-event data analysis in that the CIF of any event can be derived from the hazards. The task is to identify and estimate the counterfactual hazard of each event under the treatment condition $a \in \{0, 1\}$. A natural idea is to use inverse probability weighting (IPW). By weighting the individuals in the sample, a pseudo-population with baseline covariates distributed identically as the overall population is generated. Since the counterfactual hazard function is a conditional probability, IPW should be performed separately for incidence and at-risk probability. Denote

$$w(a; A, X) = \frac{I\{A = a\}}{e(a; X)}$$

which is the inverse of the treatment propensity score multiplied by the associated treatment indicator. We have the following identifiability result.

Theorem 1. *Under Assumptions 1–4, the counterfactual cumulative cause-specific hazard of event j is identifiable* ($a = 0, 1$, $j = 1, \ldots, J$),

$$\Lambda_j^a(t) = \int_0^t \frac{E\{w(a; A, X) dN_j(s)\}}{E\{w(a; A, X) Y(s)\}}$$

Thus, the counterfactual cumulative incidence function of event j is identifiable.

Inspired by this theorem, if the propensity score is known, an estimator for the counterfactual cumulative cause-specific hazard by IPW is given by

$$\widetilde{\Lambda}_j^a(t) = \int_0^t \frac{\sum_{i=1}^n w(a; A_i, X_i) dN_{ij}(s)}{\sum_{i=1}^n w(a; A_i, X_i) Y_i(s)} \tag{2}$$

Otherwise, if the propensity score is unknown, we plug the estimated propensity score in $w(a; A_i, X_i)$ to obtain an empirical version $\widehat{w}(a; A_i, X_i)$. The resulting estimator for the counterfactual cumulative cause-specific hazard is given by

$$\widehat{\Lambda}_j^a(t) = \int_0^t \frac{\sum_{i=1}^n \widehat{w}(a; A_i, X_i) dN_{ij}(s)}{\sum_{i=1}^n \widehat{w}(a; A_i, X_i) Y_i(s)} \tag{3}$$

The estimator for the counterfactual cumulative hazard is a step function with jumps at event times. Based on Equation (1), we obtain the following adjusted Nelson–Aalen estimator for the counterfactual cumulative incidence function of event j when the propensity score is estimated,

$$\widehat{F}_j^a(t) = \int_0^t \exp\left\{ -\sum_{k=1}^{J} \widehat{\Lambda}_k^a(s) \right\} d\widehat{\Lambda}_j^a(s),\ 0 \le t < t^* \tag{4}$$

3 | Asymptotic Properties and Inference

In this section, we derive the asymptotic variance of the adjusted Nelson–Aalen estimator. Let

$$\begin{aligned} \Psi_{1j}(t; a) &= P(T^a \le t, \Delta^a = j) = E\{w(a; A, X) N_j(t)\}, \\ \Psi_2(t; a) &= P(T^a \ge t) = E\{w(a; A, X) Y(t)\} \end{aligned}$$

then it follows from Theorem 1 that $\Lambda_j^a(t) = \int_0^t d\Psi_{1j}(s; a) / \Psi_2(s; a)$. Let

$$\psi_{1j}(t; a) = w(a; A, X) N_j(t),\ \psi_2(t; a) = w(a; A, X) Y(t)$$

We use $\mathbb{P}(\cdot)$ to denote the measure concerning the true data generating process and $\mathbb{P}_n(\cdot)$ to denote the empirical measure on the sample, so $\Psi_{1j}(t; a) = \mathbb{P}\{\psi_{1j}(t; a)\}$ and $\Psi_2(t; a) = \mathbb{P}\{\psi_2(t; a)\}$.

3.1 | When the Propensity Score is Known

We first assume that the propensity score is known. For example, individuals are assigned to treatment arms with a known probability in stratified randomized controlled trials. The oracle IPW

estimator for the counterfactual cumulative cause-specific hazard of event j is $\widetilde{\Lambda}_j^a(t) = \int_0^t \mathbb{P}_n\{d\psi_{1j}(s;a)\}/\mathbb{P}_n\{\psi_2(s;a)\}$. Let

$$M_j(t;a) = d\psi_{1j}(t;a) - \psi_2(t;a)d\Lambda_j^a(t)$$

and $\overline{M}_j(t;a) = \mathbb{P}_n\{M_j(t;a)\}$.

Lemma 1. *$\overline{M}_j(t;a)$ is a martingale with respect to the filter $\mathcal{F}_j(t;a) = \{w(a;A_i,X_i), Y_i(s) : s \le t, i = 1, \ldots, n\}$.*

Using the martingale theory, we can show that this oracle IPW estimator is unbiased and derive its variance.

Theorem 2. *Under Assumptions 1–4 and that the propensity score is known,*

$$E\{\widetilde{\Lambda}_j^a(t)\} = \Lambda_j^a(t),\ \mathrm{var}\{\widetilde{\Lambda}_j^a(t)\} = \frac{1}{n}E\left\{\int_0^t \frac{\mathbb{P}_n\{w(a;A,X)\psi_2(s;a)\}d\Lambda_j^a(s)}{[\mathbb{P}_n\{\psi_2(s;a)\}]^2}\right\}$$

The influence function (IF) of $\widetilde{\Lambda}_j^a(t)$ is

$$\mathrm{IF}\{\widetilde{\Lambda}_j^a(t)\} = \int_0^t \frac{1}{\Psi_2(s;a)}dM_j(s;a)$$

and thus

$$\sqrt{n}\{\widetilde{\Lambda}_j^a(t) - \Lambda_j^a(t)\}d\to N\left(0, E[\mathrm{IF}\{\widetilde{\Lambda}_j^a(t)\}^2]\right)$$

The finite-sample variance of $\widetilde{\Lambda}_j^a(t)$ can be unbiasedly estimated by the empirical counterpart of the variance formula. The asymptotic variance can be consistently estimated by the variance of the influence function evaluated using observed data.

3.2 | When the Propensity Score is Estimated

Suppose that the propensity score $e(a;x)$ belongs to a model indexed by θ, that is, $\{e(a;x;\theta), \theta \in \Theta\}$, where Θ is the parameter space. Furthermore, we assume $\{e(a;x;\theta), \theta \in \Theta\}$ is a Donsker class and the estimate of θ is regular and asymptotically linear (RAL), with

$$\widehat{\theta} - \theta = \mathbb{P}_n\phi + o_p(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \phi(A_i, X_i;\theta) + o_p(n^{-1/2})$$

where ϕ is the influence function of $\widehat{\theta}$. This is to say that the model for the propensity score should not be too complex [19]. For example, the propensity score can be fitted by logistic model $e(1;x;\theta) = \{1+\exp(-x'\theta)\}^{-1}$. If θ is estimated by maximum likelihood, then $\phi = [\mathbb{P}\{Xe(1;X;\theta)e(0;X;\theta)X'\}]^{-1}\{A - e(1;X;\theta)\}X$. Plugging in the estimated propensity score $\widehat{e}(a;x) = e(a;x;\widehat{\theta})$, let

$$\widehat{\psi}_{1j}(t;a) = \frac{I\{A=a\}}{e(a;X;\widehat{\theta})}N_j(t),\ \widehat{\psi}_2(t;a) = \frac{I\{A=a\}}{e(a;X;\widehat{\theta})}Y(t)$$

so the IPW estimator for the counterfactual cumulative cause-specific hazard of event j is $\widehat{\Lambda}_j^a(t) = \int_0^t \mathbb{P}_n\{d\widehat{\psi}_{1j}(s)\}/\mathbb{P}_n\{\widehat{\psi}_2(s)\}$.

Theorem 3. *Under Assumptions 1–4 and that $\widehat{\theta}$ in the propensity score model is RAL, the influence function (IF) of $\widehat{\Lambda}_j^a(t)$ is*

$$\mathrm{IF}\{\widehat{\Lambda}_j^a(t)\} = \int_0^t \frac{1}{\Psi_2(s;a)}\left[dM_j(s;a) - \mathbb{P}\left\{dM_j(s;a)\frac{\dot{e}(a;X;\theta)}{e(a;X;\theta)^2}\right\}\phi\right]$$

and thus

$$\sqrt{n}\{\widehat{\Lambda}_j^a(t) - \Lambda_j^a(t)\}d\to N\left(0, E[\mathrm{IF}\{\widehat{\Lambda}_j^a(t)\}^2]\right)$$

The proofs of the lemma and theorems are given in Supporting Information A. Compared with the influence function of the oracle estimator, there is an augmented term in the influence function of $\widehat{\Lambda}_j^a(t)$,

$$\nu_j(t;a) = -\int_0^t \frac{1}{\Psi_2(s;a)}\mathbb{P}\left\{dM_j(s;a)\frac{\dot{e}(a;X;\theta)}{e(a;X;\theta)^2}\right\}\phi$$

where $\dot{e}(a;X;\theta)$ is the derivative of $e(a;X;\theta)$ with respect to θ. The expectation of this augmented term is not zero because $\{X_i : i = 1, \ldots, n\}$ is not in the filter $\mathcal{F}_j(t;a)$. But by noticing that $\mathbb{P}_n\{M_j(t;a)\}$ is a martingale with mean zero, the expectation of this augmented term $\nu_j(t;a)$ is generally small. The variance of $\widehat{\Lambda}_j^a(t)$ can be consistently estimated by plugging the estimates into the influence function,

$$\widehat{\sigma}_{j,n}^{a,2}(t) = \frac{1}{n^2}\sum_{i=1}^n\left[\int_0^t \frac{1}{\mathbb{P}_n\{\widehat{\psi}_2(s;a)\}}\left[d\widehat{M}_{ij}(s;a) - \mathbb{P}_n\left\{d\widehat{M}_j(s;a)\frac{\dot{e}(a;X;\widehat{\theta})}{e(a;X;\widehat{\theta})^2}\right\}\phi(A_i,X_i;\widehat{\theta})\right]\right]^2 \quad (5)$$

where $\widehat{M}_{ij}(t;a) = d\widehat{\psi}_{1ij}(t;a) - \widehat{\psi}_{2i}(t;a)d\widehat{\Lambda}_j^a(t)$ is the martingale with estimates plugged in. As a comparison, the variance estimate by ignoring the uncertainty of the estimated propensity score is as follows:

$$\widetilde{\sigma}_{j,n}^{a,2}(t) = \frac{1}{n^2}\sum_{i=1}^n\left[\int_0^t \frac{1}{\mathbb{P}_n\{\widehat{\psi}_2(s;a)\}}d\widehat{M}_{ij}(s;a)\right]^2 \quad (6)$$

On the cumulative incidence scale, the influence function of $\widehat{F}_j^a(t)$ is derived by the functional delta method,

$$\mathrm{IF}\{\widehat{F}_j^a(t)\} = \int_0^t \exp\left\{-\sum_{k=1}^J \Lambda_k^a(s)\right\}d\mathrm{IF}\{\widehat{\Lambda}_j^a(s)\} - \int_0^t \sum_{k=1}^J \mathrm{IF}\{\widehat{\Lambda}_k^a(s)\}dF_j(s)$$

and thus

$$\sqrt{n}\{\widehat{F}_j^a(t) - F_j^a(t)\}d\to N\left(0, E[\mathrm{IF}\{\widehat{F}_j^a(t)\}]\right)$$

The asymptotic variance of $\widehat{F}_j^a(t)$ can be estimated by plug-in estimators. Sometimes researchers are also interested in the average treatment effect (ATE) $\tau_j(t) = F_j^1(t) - F_j^0(t)$, the difference in the counterfactual cumulative incidences of event j under the treated and control. The average treatment effect can be estimated by plug-in estimators, and the asymptotic properties can be easily obtained due to the additivity of influence functions.

4 | Simulation Studies

In this section, we conduct simulation studies to assess the finite-sample performance of the adjusted Nelson–Aalen estimator and compare confidence intervals based on Equations (5) and (6) when the propensity score is estimated.

Suppose that there are three covariates $X = (X_1, X_2, X_3)$ following the multivariate standard normal distribution. The potential failure times of the first occurrence under the treated and under the control are assumed to follow the proportional hazards model,

$$d\Lambda^1(t; x) = (t^2/5)\exp(x_2/5 + x_3/5)dt,$$
$$d\Lambda^0(t; a) = (t/3)\exp(x_2/3 - x_3/3)dt$$

from which we generate $\tilde{T}^1$ and $\tilde{T}^0$. We assume there are two competing events. The probabilities that the first type of event (which is the event of interest) occurs are

$$\begin{aligned} P(\tilde{\Delta}^1 = 1|\tilde{T}^1 = t, X = x) \\ = \text{expit}(-0.1 + 0.2x_1 + 0.2x_2 + 0.2x_3 + 0.03t), \\ P(\tilde{\Delta}^0 = 1|\tilde{T}^0 = t, X = x) \\ = \text{expit}(0.2x_1 - 0.1x_2 + 0.1x_3 + 0.05t) \end{aligned}$$

under the treated and control, respectively, where $\text{expit}(x) = 1/\{1 + \exp(-x)\}$. The censoring time is generated from a uniform distribution in [6, 12]. The propensity score is assumed to follow the logistic model,

$$P(A = 1|X = x) = \text{expit}(0.2 + 0.5x_1 - 0.5x_2)$$

To examine the empirical bias, we compare three point estimates: (1) the oracle adjusted Nelson–Aalen (NA) estimate that uses the true propensity score, (2) the adjusted Nelson–Aalen estimate that uses the estimated propensity score, and (3) the weighted Aalen–Johansen (AJ) estimate [16] by imposing the inverse of propensity score as weights in the function `survfit` in library `survival` of R [20]. We consider five methods to obtain the standard error: (1) the oracle standard error that uses the true propensity score in Theorem 2, (2) the naive standard error $\widetilde{\sigma}^a_{j,n}(t)$ that does not consider the uncertainty of the estimated propensity score, (3) the standard error $\widehat{\sigma}^a_{j,n}(t)$ that corrects for the estimated propensity score in Theorem 3, (4) the nonparametric bootstrap standard error, and (5) the standard error of the weighted Aalen-Johansen estimate calculated by `survfit`.

Next, we consider five confidence intervals: (1) the oracle confidence interval using the oracle adjusted Nelson–Aalen estimate and the oracle standard error, (2) the naive confidence interval using the adjusted Nelson–Aalen estimate and the naive standard error, (3) the corrected confidence interval using the adjusted Nelson–Aalen estimate and the corrected standard error, (4) the bootstrap confidence interval using the adjusted Nelson–Aalen estimate and the bootstrap standard error, and (5) the confidence interval using the weighted Aalen–Johansen estimate and associated standard error in `survfit`. Nominal 95% confidence intervals are constructed using the point estimate and standard error based on asymptotic normality.

Figure 1 shows 100 adjusted Nelson–Aalen estimates of the counterfactual cumulative incidence $F^a_1(t)$, $a = 1, 0$ when the sample size $n = 500$. The estimates distribute around the true value. Tables 1 and 2 show the empirical bias of the point estimates with standard deviation, mean standard error, and coverage rate of the nominal 95% confidence intervals when the sample size $n = 500$. We see negligible bias in the adjusted Nelson–Aalen estimator. Even if we use the estimated propensity score, the standard deviation (and standard error) is similar to the oracle one, indicating that the efficiency loss due to estimating the propensity score is negligible. Considering the uncertainty of the estimated propensity score, the standard error based on the

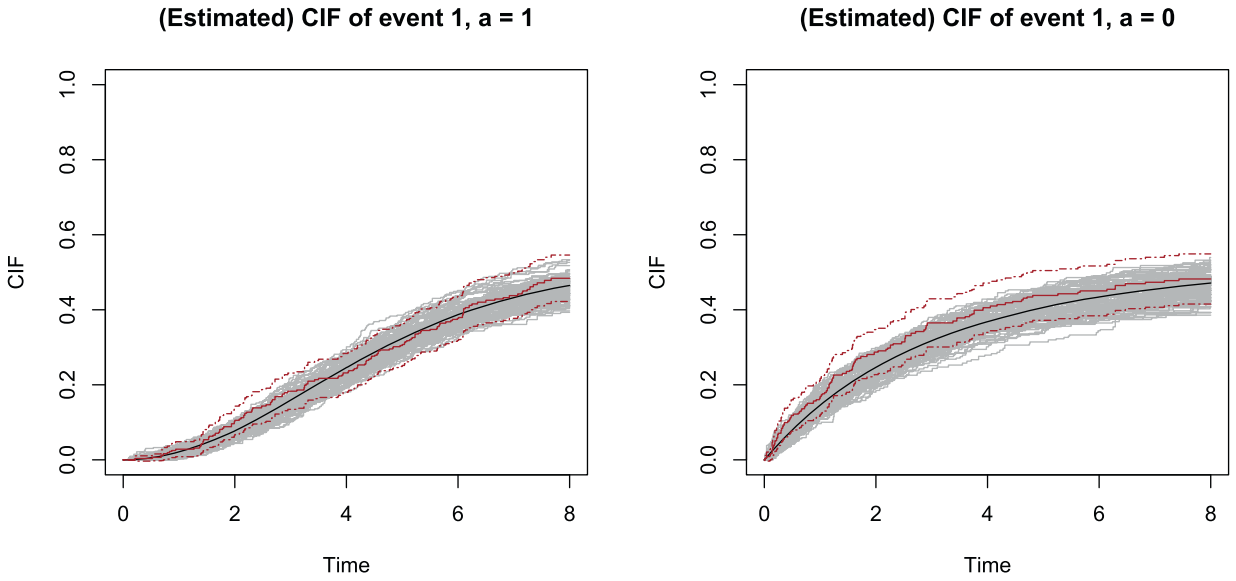

FIGURE 1 | The estimated counterfactual cumulative incidences by the adjusted Nelson–Aalen estimator. The black line is the true value, and the gray lines are the estimates (we draw 100 lines). The solid brown line is a single estimate, and the dashed brown lines are the upper and lower bounds of the 95% confidence interval associated with this estimate.

TABLE 1 | Bias, standard deviation, mean standard error, and confidence interval coverage rate of some estimators for the counterfactual cumulative incidence function $F_1^1(t)$ when the sample size $n = 500$.

Time	1	2	3	4	5	6	7	8
Bias								
Oracle	−0.000	−0.001	−0.001	−0.000	−0.000	−0.001	−0.002	−0.003
Adjusted NA	−0.000	−0.001	−0.001	−0.000	−0.001	−0.001	−0.002	−0.003
Weighted AJ	−0.000	−0.000	0.000	0.001	0.001	0.002	0.001	0.001
Standard deviation (SD)								
Oracle	0.009	0.017	0.024	0.028	0.029	0.029	0.030	0.032
Adjusted NA	0.009	0.017	0.024	0.028	0.029	0.030	0.030	0.032
Weighted AJ	0.009	0.018	0.024	0.028	0.029	0.030	0.031	0.032
Standard error (SE)								
Oracle	0.009	0.017	0.023	0.027	0.030	0.031	0.031	0.032
Naive	0.009	0.017	0.023	0.028	0.030	0.031	0.032	0.032
Corrected	0.009	0.019	0.027	0.031	0.033	0.035	0.035	0.035
Bootstrap	0.009	0.017	0.023	0.027	0.029	0.030	0.031	0.031
Weighted AJ	0.009	0.017	0.024	0.028	0.030	0.031	0.032	0.033
Coverage rate								
Oracle	0.886	0.934	0.950	0.941	0.955	0.955	0.957	0.955
Naive	0.887	0.935	0.946	0.946	0.957	0.957	0.966	0.952
Corrected	0.889	0.942	0.959	0.957	0.967	0.969	0.973	0.968
Bootstrap	0.885	0.933	0.939	0.936	0.952	0.953	0.958	0.948
Weighted AJ	0.887	0.938	0.950	0.946	0.958	0.963	0.965	0.958

TABLE 2 | Bias, standard deviation, mean standard error, and confidence interval coverage rate of some estimators for the counterfactual cumulative incidence function $F_1^0(t)$ when the sample size $n = 500$.

Time	1	2	3	4	5	6	7	8
Bias								
Oracle	−0.001	−0.002	−0.003	−0.003	−0.003	−0.004	−0.005	−0.005
Adjusted NA	−0.001	−0.002	−0.003	−0.003	−0.003	−0.004	−0.005	−0.005
Weighted AJ	0.000	−0.000	−0.000	0.000	0.000	0.000	−0.000	−0.000
Standard deviation (SD)								
Oracle	0.025	0.031	0.034	0.034	0.035	0.035	0.036	0.037
Adjusted NA	0.025	0.031	0.033	0.034	0.035	0.036	0.037	0.037
Weighted AJ	0.025	0.031	0.034	0.034	0.035	0.036	0.037	0.037
Standard error (SE)								
Oracle	0.024	0.030	0.033	0.034	0.034	0.035	0.035	0.035
Naive	0.025	0.030	0.033	0.034	0.035	0.035	0.035	0.035
Corrected	0.025	0.031	0.033	0.035	0.035	0.036	0.036	0.036
Bootstrap	0.024	0.030	0.032	0.033	0.034	0.034	0.034	0.035
Weighted AJ	0.025	0.031	0.033	0.035	0.035	0.036	0.036	0.036
Coverage rate								
Oracle	0.937	0.942	0.931	0.949	0.937	0.942	0.938	0.941
Naive	0.932	0.946	0.932	0.947	0.941	0.943	0.934	0.941
Corrected	0.941	0.951	0.935	0.947	0.943	0.950	0.942	0.945
Bootstrap	0.934	0.942	0.927	0.940	0.940	0.939	0.934	0.935
Weighted AJ	0.939	0.951	0.940	0.948	0.947	0.951	0.941	0.944

influence function in Theorem 3 is slightly larger than the naive one, and hence the coverage rate based on the corrected standard error is larger. The weighted Aalen–Johansen estimator from the R function `survfit` gives a similar standard error to the naive adjusted Nelson–Aalen estimator, and these two types of estimators are actually asymptotically equivalent with slightly different finite-sample performances [21, 22]. The most interesting finding is that omitting the variation of the estimated propensity score does not lead to much bias of the standard error. In Supporting Information C, we provide more simulation results, including the setting where the sample size is $n = 2000$ and a sensitivity analysis where the propensity score model is misspecified.

5 | Application to Allogeneic Stem Cell Transplantation Data

Allogeneic stem cell transplantation is a widely applied therapy to treat acute lymphoblastic leukemia (ALL), including two sorts of transplant modalities: human leukocyte antigens matched sibling donor transplantation (MSDT) and haploidentical stem cell transplantation from family (Haplo-SCT). MSDT has long been regarded as the first choice of transplantation because MSDT leads to lower transplant-related mortality, also known as nonrelapse mortality (NRM) [23]. In recent years, some benefits of Haplo-SCT have been noticed that patients with positive pretransplantation minimum residual disease (MRD) undergoing Haplo-SCT have better prognosis in relapse, and hence lower relapse-related mortality [24]. It is interesting to investigate whether Haplo-SCT can be an alternative to MSDT since the former is much more accessible. We adopt the relapse as the primary event and nonrelapse mortality as the competing event.

A total of $n = 303$ patients with positive MRD undergoing allogeneic stem cell transplantation at Peking University People's Hospital in China from 2009 to 2017 were included in our study [25]. Among these patients, 65 received MSDT ($A = 0$) and 238 received Haplo-SCT ($A = 1$). There is no specific consideration to prefer Haplo-SCT over MSDT whenever MSDT is accessible [24], so we expect ignorability. Four baseline covariates are considered: age, sex (male or female), diagnosis (T-cell ALL [T-ALL] or B-cell ALL [B-ALL]) and complete remission status (after 1 cycle [CR1] or more than 1 cycle [CR> 1]). As found in previous literature, these covariates are risk factors associated with relapse and mortality. The time origin is the time of receiving transplantation (either MSDT or Haplo-SCT). The mean follow-up time is 1336 days, and the maximum follow-up time is 4106 days. In the MSDT group, 47.7% patients were observed to encounter relapse and 9.2% NRM. In the Haplo-SCT group, 29.0% patients were observed to encounter relapse and 11.8% NRM. Summary statistics are presented in Table 3.

The propensity score is fitted by logistic regression. The upper-left panel in Figure 2 shows the estimated counterfactual cumulative incidence function of relapse. The relapse rate is higher in the MSDT group than in the Haplo-SCT group. The 95% confidence intervals are displayed in dashed lines. We plot the confidence interval that ignores the uncertainty of the estimated propensity score and the confidence interval that corrects for the estimated propensity score. The numerical values of these two types of confidence intervals are very similar, and we cannot visually tell a difference between these two types of confidence intervals. The upper-right panel of Figure 2 presents the average treatment effect (ATE) of Haplo-SCT on the relapse rate compared to MSDT, defined as the difference in counterfactual cumulative incidences. The 95% confidence interval calculated based on the influence functions of the estimated CIFs is displayed in dashed lines. Negative values of the ATE indicate that Haplo-SCT leads to a lower relapse rate than MSDT. The bottom-left panel shows the estimated cumulative incidence function of NRM and the bottom-right panel shows the estimated average treatment effect. The NRM rate is slightly higher in the Haplo-SCT group than in the MSDT group. We list the estimated treatment effects and 95% confidence intervals at some time points in Table 4.

In summary, we find that Haplo-SCT significantly reduces the risk of relapse compared to MSDT. Haplo-SCT can be used as the better transplant modality for MRD positive patients. Since Haplo-SCT is more accessible than MSDT, it is promising that Haplo-SCT be considered as the first choice of allogeneic stem

TABLE 3 | Summary statistics in the stem cell transplantation data, stratified by the treatment groups. We list the mean and standard deviation (SD) for continuous variables, and the count and proportion for binary variables.

	Haplo-SCT ($A = 1$)		MSDT ($A = 0$)	
	Mean/Count	**SD/Proportion**	**Mean/Count**	**SD/Proportion**
Baseline covariates				
Age	26.7	(12.2)	35.0	(13.1)
Sex: Male	89	37.4%	27	41.5%
Complete Remission: CR1	54	22.7%	10	15.4%
Diagnosis: T-ALL	38	16.0%	3	4.6%
Observed events of first occurrence				
Relapse	69	29.0%	31	47.7%
Time to relapse (days)	371.7	(406.2)	420.8	(369.8)
NRM	28	11.8%	6	9.2%
Time to NRM (days)	192.2	(192.0)	207.5	(95.2)

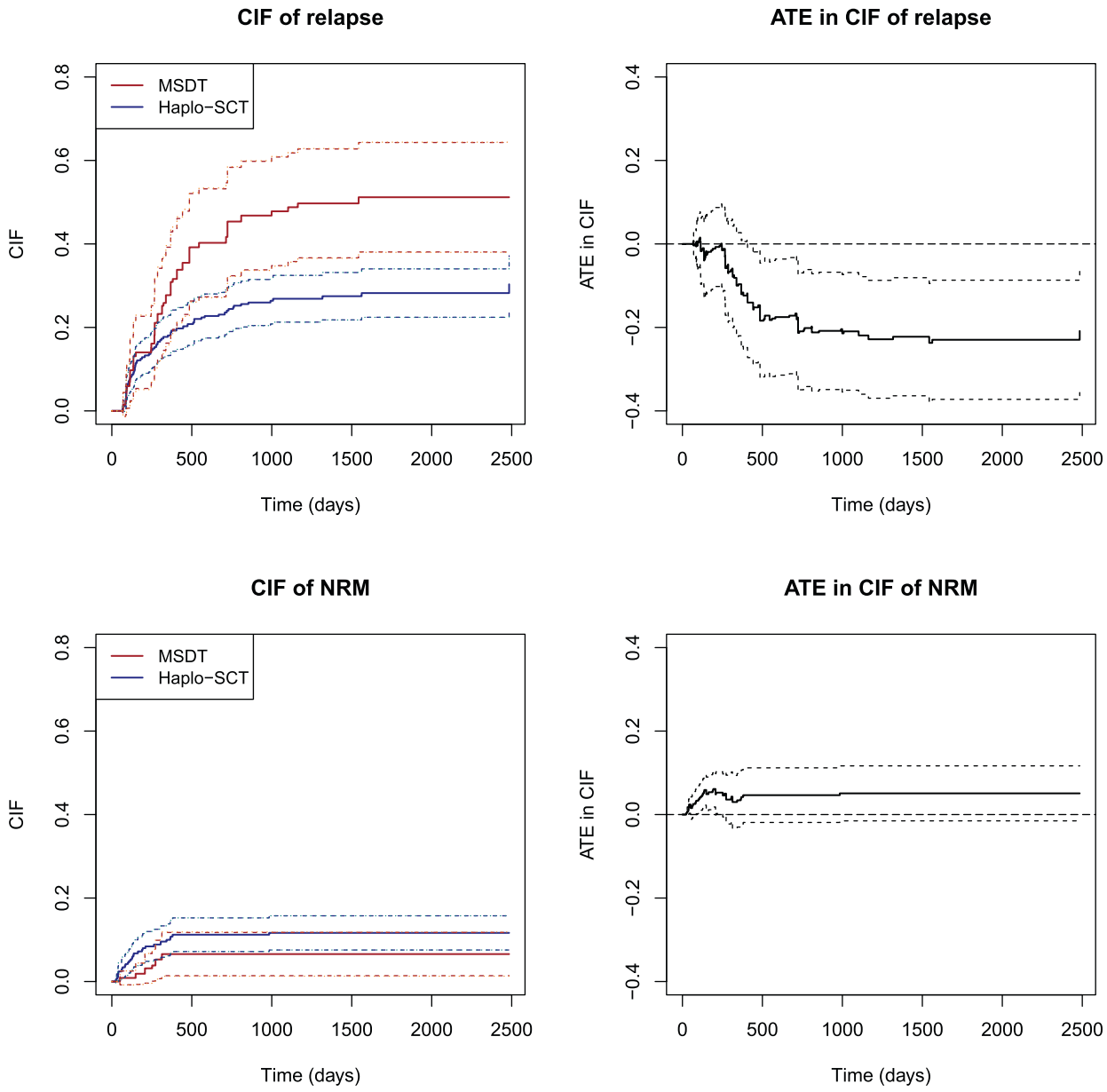

FIGURE 2 | Left: The estimated counterfactual cumulative incidence functions of relapse and nonrelapse mortality (NRM). The brown line is MSDT and the blue line is Haplo-SCT. The brown and blue dashed lines are the 95% confidence intervals by correcting for the estimated propensity score; the orange and cyan lines are the 95% confidence intervals ignoring the uncertainty of the estimated propensity score. Right: the average treatment effect (ATE) on the incidence scale. Negative values indicate that Haplo-SCT leads to a lower relapse (NRM) rate than MSDT.

TABLE 4 | Average treatment effects (ATEs) on the counterfactual cumulative incidence functions of relapse and NRM with 95% confidence intervals (CIs) at some time points.

Years after	Relapse			NRM		
transplantation	ATE	95% CI	*P*-value	ATE	95% CI	*P*-value
0.5	−0.015	(−0.109, 0.079)	0.752	0.054	(0.012, 0.095)	0.012
1.0	−0.096	(−0.220, 0.027)	0.126	0.038	(−0.026, 0.103)	0.244
2.0	−0.209	(−0.349, −0.069)	0.003	0.046	(−0.019, 0.112)	0.164
3.0	−0.209	(−0.350, −0.069)	0.004	0.051	(−0.015, 0.117)	0.131
5.0	−0.230	(−0.372, −0.087)	0.002	0.051	(−0.015, 0.117)	0.131

cell transplantation as long as paying more care to prevent transplant-related mortality.

6 | Discussion

Inverse probability weighting has been well applied in causal inference due to its simplicity and interpretability. However, the estimand (cumulative incidence function) is time-varying for time-to-event outcomes. We may employ IPW in two ways: the first is to directly estimate the incidence by weighting the observed event counting process by the inverse of propensity score and uncensored probability, and the second is to estimate the population-level hazard by weighting both the event counting process and at-risk process. In this article, we follow the second approach.

There are a few advantages of estimating the cumulative incidence through transforming hazards over directly targeting the cumulative incidence. First, only one model (the treatment propensity score) is required. We do not need to estimate the censoring probability as long as the censoring is completely noninformative or estimate the conditional hazards with complicated post-treatment (time-varying) covariates. Second, the additional variance resulted by the uncertainty of the estimated propensity score is negligible since the extra term in the influence function of $\widehat{\Lambda}_j^a(t)$ is a weighted martingale, although the expectation may not be exactly zero.

To conclude, we point out five possible extensions for the proposed method. First, there are alternatives on the estimation of the propensity score. Different parametric models like logistic regression and probit regression can be used. Since IPW essentially uses the balancing property of the propensity score, we can adapt the loss functions to obtain the covariates balancing propensity score [26]. As long as the estimated propensity score has a known form of influence function, the asymptotic variance of the adjusted Nelson–Aalen estimator can be corrected.

Second, estimating the hazard is potentially helpful for mediation analysis. Within the interventional effects framework, suppose we can draw the event counting process from some reference distribution (such as according to the counterfactual cause-specific hazard under control). Under some additional assumptions, we can identify and estimate the counterfactual cumulative incidence of each event under such an intervention. This allows one to study the direct and indirect effects on a single event. Testing the treatment effect can be reduced to an implication of testing the counterfactual cause-specific hazards, which can be achieved by logrank tests. However, the uncertainty of the estimated score should be accounted for in the tests.

Third, the competing risks framework can be extended to multistate models. Typical assumptions in multistate models are Markovness or semi-Markovness [27, 28]. The former says that the transition hazard from one state to another only relies on the time since the origin rather than the history. In contrast, the latter says that the transition hazard only depends on how long it has passed since reaching the last state. With completely noninformative censoring, the transition hazards can be consistently estimated by IPW, and the cumulative incidence of any state can then be estimated after some standard derivation. The asymptotic properties of the estimated incidence can be established using the functional delta method.

Fourth, the assumptions in our framework may be relaxed. If the censoring is not completely random but depends on observed covariates, the counterfactual hazard is still identifiable but more complicated. In addition to the propensity score, the censoring probability should also be adjusted to reflect the overall population. Informative censoring induces a biased selection issue, where the at-risk individuals may have different underlying features in the real world (subject to censoring) and counterfactual world (not subject to censoring). Therefore, the counting processes should be additionally weighted by the conditional uncensored probability, which introduces another source of variance. Suppose the estimated conditional uncensored probability is RAL (e.g., by Cox regression). In that case, we can imitate the same strategy considered in this article to derive the asymptotic variance of the estimated counterfactual CIF.

Fifth, the efficiency can be improved by incorporating models for the failure times. We can derive the efficient influence function (EIF) for the counterfactual CIF under conditionally random censoring (which is weaker than completely random censoring), which involves the propensity score, censoring probability and cause-specific hazards for all events [29–31]. The explicit form of the EIF is given in the Supporting Information B. This estimator based on EIF does not have the adjusted Nelson–Aalen form. In practice, the propensity score is fitted at baseline, whereas the censoring probability and cause-specific hazards are fitted using post-treatment data. Modeling the univariate propensity score is computationally easier by regression and provides desirable asymptotic properties. It is challenging to correctly specify the hazards if the dependence of events is complex or if there are time-varying covariates. Misspecification of working models may lead to bias and inconsistent variance estimation.

Acknowledgments

We thank Dr. Wang Miao (Peking University) for comments. We also thank Dr. Yingjun Chang, Dr. Leqing Cao, and Dr. Yuewen Wang (Peking University People's Hospital) for cleaning the data.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that supports the findings of this study are available in the Supporting Information of this article.

References

1. R. L. Prentice, J. D. Kalbfleisch, A. Peterson Jr, N. Flournoy, V. Farewell, and N. Breslow, "The Analysis of Failure Times in the Presence of Competing Risks," *Biometrics* 34, no. 4 (1978): 541–554.

2. B. Lau, S. R. Cole, and S. J. Gange, "Competing Risk Regression Models for Epidemiologic Data," *American Journal of Epidemiology* 170, no. 2 (2009): 244–256.

3. B. R. Logan, J. P. Klein, and M. J. Zhang, "Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation," *Biometrics* 64, no. 3 (2008): 733–740.

4. H. C. Rytgaard, T. A. Gerds, and M. J. van der Laan, "Continuous-Time Targeted Minimum Loss-Based Estimation of Intervention-Specific Mean Outcomes," *Annals of Statistics* 50, no. 5 (2022): 2469–2491.

5. E. L. Kaplan and P. Meier, "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association* 53, no. 282 (1958): 457–481.

6. W. Nelson, "Theory and Applications of Hazard Plotting for Censored Failure Data," *Technometrics* 14, no. 4 (1972): 945–966.

7. O. O. Aalen, "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics* 6, no. 4 (1978): 701–726.

8. P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, no. 1 (1983): 41–55.

9. J. M. Robins, M. A. Hernan, and B. Brumback, "Marginal Structural Models and Causal Inference in Epidemiology," *Epidemiology* 11, no. 5 (2000): 550–560.

10. A. Winnett and P. Sasieni, "Adjusted Nelson–Aalen Estimates With Retrospective Matching," *Journal of the American Statistical Association* 97, no. 457 (2002): 245–256.

11. J. Xie and C. Liu, "Adjusted Kaplan–Meier Estimator and Log-Rank Test With Inverse Probability of Treatment Weighting for Survival Data," *Statistics in Medicine* 24, no. 20 (2005): 3089–3110.

12. P. C. Austin, "The Use of Propensity Score Methods With Survival or Time-To-Event Outcomes: Reporting Measures of Effect Similar to Those Used in Randomized Experiments," *Statistics in Medicine* 33, no. 7 (2014): 1242–1258.

13. H. Mao, L. Li, W. Yang, and Y. Shen, "On the Propensity Score Weighting Analysis With Survival Outcome: Estimands, Estimation, and Inference," *Statistics in Medicine* 37, no. 26 (2018): 3745–3763.

14. G. Hu and F. Huffer, "Modified Kaplan–Meier Estimator and Nelson–Aalen Estimator With Geographical Weighting for Survival Data," *Geographical Analysis* 52, no. 1 (2020): 28–48.

15. B. Vakulenko-Lagun, C. Magdamo, M. L. Charpignon, B. Zheng, M. W. Albers, and S. Das, "causalCmprsk: An R Package for Nonparametric and Cox-Based Estimation of Average Treatment Effects in Competing Risks Data," *Computer Methods and Programs in Biomedicine* 242 (2023): 107819.

16. O. O. Aalen and S. Johansen, "An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations," *Scandinavian Journal of Statistics* 5, no. 3 (1978): 141–150.

17. S. R. Cole and M. A. Hernán, "Adjusted Survival Curves With Inverse Probability Weights," *Computer Methods and Programs in Biomedicine* 75, no. 1 (2004): 45–49.

18. K. Hirano, G. W. Imbens, and G. Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71, no. 4 (2003): 1161–1189.

19. A. W. van der Vaart and J. A. Wellner, "Empirical Processes," in *Weak Convergence and Empirical Processes: With Applications to Statistics* (Springer, 2023), 127–384.

20. T. M. Therneau, "A Package for Survival Analysis in R," 2024 R package version 3.8–3.

21. D. Luo and S. C. Saunders, "Bias and Mean-Square Error for the Kaplan–Meier and Nelson–Aalen Estimators," *Journal of Nonparametric Statistics* 3, no. 1 (1993): 37–51.

22. E. Colosimo, F. Ferreira, M. Oliveira, and C. Sousa, "Empirical Comparisons Between Kaplan–Meier and Nelson–Aalen Survival Function Estimators," *Journal of Statistical Computation and Simulation* 72, no. 4 (2002): 299–308.

23. C. G. Kanakry, E. J. Fuchs, and L. Luznik, "Modern Approaches to HLA-Haploidentical Blood or Marrow Transplantation," *Nature Reviews Clinical Oncology* 13, no. 1 (2016): 10–24.

24. Y. J. Chang, Y. Wang, L. P. Xu, et al., "Haploidentical Donor Is Preferred Over Matched Sibling Donor for Pre-Transplantation MRD Positive ALL: A Phase 3 Genetically Randomized Study," *Journal of Hematology & Oncology* 13, no. 1 (2020): 1–13, https://doi.org/10.1186/s13045-020-00860-y.

25. R. Ma, L. P. Xu, X. H. Zhang, et al., "An Integrative Scoring System Mainly Based on Quantitative Dynamics of Minimal/Measurable Residual Disease for Relapse Prediction in Patients With Acute Lymphoblastic Leukemia," *Journal of Clinical Oncology* 39, no. 15_suppl (2021).

26. K. Imai and M. Ratkovic, "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 76, no. 1 (2014): 243–263.

27. D. Commenges, P. Joly, A. Gégout-Petit, and B. Liquet, "Choice Between Semi-Parametric Estimators of Markov and Non-Markov Multi-State Models From Coarsened Observations," *Scandinavian Journal of Statistics* 34, no. 1 (2007): 33–52.

28. A. Asanjarani, B. Liquet, and Y. Nazarathy, "Estimation of Semi-Markov Multi-State Models: A Comparison of the Sojourn Times and Transition Intensities Approaches," *International Journal of Biostatistics* 18, no. 1 (2022): 243–262.

29. M. Zhang and D. E. Schaubel, "Contrasting Treatment-Specific Survival Using Double-Robust Estimators," *Statistics in Medicine* 31, no. 30 (2012): 4255–4268.

30. T. Martinussen and M. J. Stensrud, "Estimation of Separable Direct and Indirect Effects in Continuous Time," *Biometrics* 79, no. 1 (2023): 127–139.

31. H. C. Rytgaard and M. J. van der Laan, "Targeted Maximum Likelihood Estimation for Causal Inference in Survival and Competing Risks Analysis," *Lifetime Data Analysis* 30, no. 1 (2024): 4–33.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.