Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE OPEN ACCESS

Inverse Probability of Treatment Weighting Using the Propensity Score With Competing Risks in Survival Analysis

Peter C. Austin[1,2,3] | Jason P. Fine[4]

[1]ICES, Toronto, Ontario, Canada | [2]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada | [3]Schulich Heart Research Program, Sunnybrook Research Institute, Toronto, Ontario, Canada | [4]Department of Statistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Correspondence: Peter C. Austin (peter.austin@ices.on.ca)

Received: 27 May 2024 | **Revised:** 9 January 2025 | **Accepted:** 13 January 2025

Funding: This work was supported by the Canadian Institutes of Health Research (PJT 183902).

Keywords: competing risk | cumulative incidence function | inverse probability of treatment weighting | propensity score | survival analysis

ABSTRACT

Inverse probability of treatment weighting (IPTW) using the propensity score allows estimation of the effect of treatment in observational studies. We had three objectives: first, to describe methods for using IPTW to estimate the effects of treatments in settings with competing risks; second, to illustrate the application of these methods using empirical analyses; and third, to conduct Monte Carlo simulations to evaluate the relative performance of three methods for estimating time-specific risk differences and time-specific relative risks in settings with competing risks. In doing so, we provide guidance to applied biostatisticians and clinical investigators on the use of IPTW in settings with competing risks. We examined three estimators of time-specific risk differences and relative risks: the weighted Aalen–Johansen estimator, an estimator that combines IPTW with inverse probability of censoring weights (IPTW-IPCWs), and a double-robust augmented IPTW estimator combined with IPCW (AIPTW-IPCW). The design of our simulations reflected clinically realistic scenarios. Our simulations found that all three estimators tended to result in unbiased estimations of time-specific risk differences and time-specific relative risks. However, the weighted Aalen–Johansen estimator and the AIPTW-IPCW estimator tended to result in estimates with greater precision compared to the IPTW-IPCW estimator. In our empirical analyses, we illustrated the application of these methods by estimating the effect of statin prescribing on the risk of subsequent cardiovascular death in patients discharged from the hospital with a diagnosis of acute myocardial infarction.

1 | Background

Investigators are increasingly using observational studies to estimate the effects of treatments, exposures, and interventions. However, a consequence of the lack of random treatment assignment is that treated subjects often differ systematically at baseline from control subjects. Because of the confounding that occurs when the distribution of baseline characteristics differs between treated and control subjects, outcomes cannot be compared directly between treated and control subjects. Instead, statistical methods must be used to remove the effects of the confounding due to measured variables. Statistical methods based on the propensity score are being used with increasing frequency in observational studies examining the effect of treatments. The propensity score is defined as a subject's probability of receiving the treatment of interest conditional on measured baseline covariates [1, 2]. There are four ways of using the propensity score: matching on the propensity score, inverse probability of treatment weighting (IPTW) using the propensity score, stratification on the propensity score, and covariate adjustment using

the propensity score. Of these four approaches, matching and IPTW tend to have superior performance compared to the other two approaches [3, 4].

Survival or time-to-event outcomes occur frequently in biomedical and epidemiological research [5]. Several papers have examined the use of propensity score methods with time-to-event outcomes [4, 6–9]. The focus of these papers was on the application of propensity score methods in settings with a single cause of failure (e.g., all-cause mortality). Competing risks are events whose occurrence precludes the occurrence of the primary event of interest [10–13]. If the primary event of interest is time to death due to cardiovascular causes, then death due to non-cardiovascular causes would serve as a competing risk, as subjects who die of non-cardiovascular causes are no longer at risk of death due to cardiovascular causes. In general, all nonfatal outcomes and all cause-specific mortality outcomes are subject to competing risks. Despite the frequency with which competing risks are present in medical and epidemiological research, only one paper has systematically examined the use of matching on the propensity score in the presence of competing risks [14].

Despite the relative popularity of IPTW and its good performance in settings without competing risks, only a few papers have described methods for using IPTW in settings with competing risks. Cole and colleagues used inverse probability weights with the Aalen–Johansen estimator of the cumulative incidence function (CIF) [15, 16]. By comparing the weighted CIFs for treated and control subjects at specific time points, they estimated the difference in the risk of the outcome or the relative risk of the outcome at these specific times. We will subsequently refer to this method as a weighted Aalen–Johansen estimator of the risk difference or relative risk. Ozenne and colleagues developed two estimators for time-specific risk differences in settings with competing risks [17]. The first combined IPTW with inverse probability of censoring weights (IPCWs), while the second was a doubly robust estimator that used both IPTW and IPCW. Both these methods require specifying a distribution for censoring times. A doubly robust estimator is an estimator that combines both an adjusted regression model for the outcome and a regression model for treatment selection. The estimate will be unbiased if at least one of these two regression models is specified correctly. We refer to these two methods as IPTW-IPCW and AIPTW-IPCW, where AIPTW refers to augmented inverse probability of treatment weighting.

The objective of the paper was three-fold: first, to describe methods for using IPTW to estimate the effects of treatments in settings with competing risks. These methods are based on estimating the CIF within each treatment group and then obtaining time-specific estimates of risk in treated and control subjects separately. Using these estimates, one can obtain time-specific estimates of risk differences and relative risks. Second, to illustrate the application of these methods using empirical analyses. Third, to conduct Monte Carlo simulations to evaluate the relative performance of different methods for estimating time-specific risk differences and relative risks in settings with competing risks. In doing so, we aim to provide guidance to applied biostatisticians and clinical investigators on the use of IPTW using the propensity score in settings in which competing risks are present. We restrict our attention in the setting of a binary point exposure that is applied at baseline. We do not consider scenarios in which covariates and treatment vary over time.

The paper is structured as follows: in Section 2, we provide a brief discussion of target estimands in the setting of competing risks. In Section 3, we describe statistical methods for estimating the effect of treatment in settings with competing risks when using IPTW. In Section 4, we present a case study to illustrate the application of these methods. In Section 5, we describe the design of a series of Monte Carlo simulations that were used to compare the relative performance of methods for estimating risk differences and relative risks at specific time points when using IPTW. In Section 6, we report the results of these simulations. Finally, in Section 7, we summarize our findings and discuss them in the context of the existing literature.

2 | Target Estimands in Settings With Competing Risks

We summarize the target estimands proposed by Cole and colleagues for settings with competing risks [15]. Their framework for defining target estimands uses the potential outcomes framework, which was initially introduced by Rubin [18]. We assume a primary outcome of interest (e.g., death due to cardiovascular causes) and a single competing event (e.g., death due to non-cardiovascular causes). The cumulative incidence of the primary outcome is defined as: $F_1(t) = \Pr\left(T_i < t, D_i = 1\right)$, where T_i denotes the time to either cardiovascular death or non-cardiovascular death for the ith subject, and D_i is an event type indicator denoting the type of event that occurred: $D_i = 1$ for cardiovascular death and $D_i = 2$ for non-cardiovascular death. Let Z be a binary variable denoting treatment status: $Z = 0$ for control and $Z = 1$ for the treatment of interest. Let $T_i(0)$ and $T_i(1)$ denote the potential times to cardiovascular death or non-cardiovascular death for the ith subject under control and treatment, respectively. Similarly, let $D_i(0)$ and $D_i(1)$ denote the potential event type indicators for this subject under control and treatment, respectively.

The potential cumulative incidences of cardiovascular death at time t under control and treatment are defined as: $F_1^0(t) = \Pr\left(T_i(0) < t, D_i(0) = 1\right)$ and $F_1^1(t) = \Pr\left(T_i(1) < t, D_i(1) = 1\right)$, respectively. The two estimands proposed by Cole and colleagues are the cardiovascular death risk difference at time t: $\mathrm{RD}(t) = F_1^1(t) - F_1^0(t)$, and the cardiovascular death risk ratio (or relative risk): $\mathrm{RR}(t) = F_1^1(t)/F_1^0(t)$. These estimands can be estimated at clinically meaningful times t.

Young and colleagues [19] describe different estimands in settings with competing risks. For a given treatment or intervention, they contrast the total effect with the direct effect. The total effect is the estimand proposed by Cole and colleagues, as defined above. In contrast, the direct effect is the effect of the treatment in a hypothetical setting in which competing events have been eliminated. Thus, the direct effect represents the effect of treatment on the outcome of interest that is not mediated by competing events. They also briefly describe another estimand, the survivor average causal effect, which is defined as the average effect of the treatment on the outcome of interest in those subjects who would never experience the competing event. Stensrud

and colleagues decompose the total effect into the sum of the separable direct effect and the separate indirect effect [20, 21]. The separable direct effect is equivalent to Young's direct effect, whereas the separable indirect effect is the effect of the treatment on the primary event of interest that is mediated through its effect on the competing event.

Young and colleagues suggest that another estimand in settings with competing risks is the effect of treatment on a composite outcome consisting of all the event types. Although this approach simplifies both the conceptual framework and the subsequent analyses, it shifts the scientific question from the effect of treatment on a specific outcome to its effect on a different outcome, of which the given outcome is just one component.

Finally, hazard ratios are a common measure of association in clinical and epidemiological research. There is a lack of consensus in the methodological literature regarding the appropriateness of using hazard ratios as causal measures of effect. Young and colleagues suggested that, in general, comparisons of hazard rates between treatment groups do not have a causal interpretation [19]. This view contrasts with that of Fay and Li, who suggest that the population-level hazard ratio (also known as the population-averaged hazard ratio or the marginal hazard ratio) has a causal interpretation [22]. Fay and Li note that one must distinguish between population-level and individual-level interpretation of hazard ratios (due to the non-collapsibility of the hazard ratio, these two quantities will not, in general, coincide [23]). They suggest that the population-level hazard ratio is a useful estimand but it must be interpreted appropriately.

In what follows we focus on two estimands: the total effect (i.e., the estimand described by Cole and colleagues) and the population-average hazard ratio. We focus on the total effect for two reasons: (i) it appears to be used more frequently than the direct effect and the survivor average causal effect; (ii) it does not require making assumptions that the competing risk can be eliminated for all individuals or for some individuals. We focus on the population-averaged hazard ratio because of the pervasive use of hazard ratios in clinical and epidemiological settings with competing risks.

3 | Statistical Methods for IPTW Using the Propensity Score in the Presence of Competing Risks

Several studies have examined the application of propensity score methods to settings with time-to-event or survival outcomes [4, 6–8, 24]. One study examined the performance of propensity score matching when competing risks are present [14]. Cole and colleagues suggested that, when competing risks are present, the weighted Aalen–Johansen estimator can be used to estimate risk differences and relative risks at specific times t [15, 16]. Ozenne developed two estimators for the risk difference when competing risks are present, both of which use IPTW and require specifying a model for the censoring distribution [17].

Dorn suggested that, when designing and analyzing a retrospective study, one asks the question "How would the study be conducted if it were possible to do it by controlled experimentation" [25]. This suggests that the analyses conducted when using propensity score weighting should mirror the analyses that would be done in an RCT with the same intervention and outcome. Several clinical commentators have suggested that medical decision-making is better informed by absolute measures of effect than by relative measures of effect [26–28], whereas others have suggested that the reporting of relative measures of effect should be complemented by the reporting of absolute measures of effect [29, 30]. The BMJ requires that the absolute risk reduction and the number needed to treat (NNT, computed as the reciprocal of the absolute risk difference) be reported for any RCT with a binary outcome [31]. This suggests that, in studies with randomized trials with time-to-event outcomes, investigators should report both absolute and relative effects. Absolute effects can be estimated by comparing survival curves between treatment groups. From these, the NNT can be computed at any duration of follow-up [32]. The relative effect of treatment can be determined from a Cox proportional hazards model in which the hazard of the event is regressed on treatment status. The resultant measure of effect is the hazard ratio, which quantifies the relative change in the hazard of the event due to treatment. Alternatively, one can estimate relative risks at specific durations of follow-up, keeping in mind that hazard ratios and relative risk do not, in general, coincide [33]. Furthermore, even when the hazard ratio is constant over time, the time-specific relative risks can differ across time points.

In this section, we describe how IPTW using the propensity score can be used in settings with competing risks. In particular, we describe how both absolute and relative measures of effect can be estimated when using IPTW in settings with competing risks.

3.1 | Background and Notation

Let Z be a binary variable denoting treatment status ($Z = 0$ for control vs. $Z = 1$ for treated), and let $\mathbf{X}$ denote a vector of observed baseline covariates. Then, the propensity score is defined as: $e(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$ [1]. The propensity score is typically estimated using a logistic regression model in which treatment status is regressed on the vector of observed covariates. Previous research has demonstrated that, in practice, variables that either confound the treatment–outcome relationship or are prognostic of the outcome should be included in the propensity score model [34].

Inverse probability of treatment weights are defined as: $w = \frac{Z}{e(\mathbf{X})} + \frac{1-Z}{1-e(\mathbf{X})}$ [35]. One can also define stabilized inverse probability of treatment weights as: $w = \Pr(Z = 1)\frac{Z}{e(\mathbf{X})} + \Pr(Z = 0)\frac{1-Z}{1-e(\mathbf{X})}$ [36, 37]. Stabilized weights are intended to improve performance. Their use results in narrower confidence intervals compared to the variance inflation that occurs due to treated subjects having very low propensity scores or control subjects having very high propensity scores.

3.2 | Estimating the Effect of Treatment on Cause-Specific Risk of the Outcome Within Specified Duration of Time When Using IPTW: Absolute and Relative Measures of Effect

As noted in Section 3.1, the absolute risk difference and the corresponding NNT are key quantities when reporting the

results of RCTs. Cole and colleagues suggested using a weighted Aalen–Johansen estimator to estimate CIFs under treatment and control [15, 16]. They proposed using inverse probability weights, which are the product of inverse probability of treatment weights (to account for nonrandom treatment assignment) and IPCWs (to account for nonrandom censoring) From the weighted CIFs, one can extract the risk of the outcome of interest under treatment and control at a specified time t, and compute the corresponding risk difference. They proposed using bootstrapping to compute confidence intervals for the risk difference.

Ozenne and colleagues developed an estimator for the risk difference and its standard error obtained from an inverse probability weighted estimator of the CIF, which requires specifying a model for the censoring distribution [17]. They also developed a doubly robust estimator for the risk difference when the ATE is the target estimand, along with a variance estimator. This estimator requires specifying three regression models: (i) an outcome regression model relating the cause-specific hazard of the outcomes to a set of subject characteristics, including treatment status; (ii) a treatment regression model relating the log-odds of treatment status to a set of subject characteristics; and (iii) a model for the censoring distribution (which can be a null Cox proportional hazards model). Once the risk difference and its associated confidence interval have been constructed, the associated NNT and its confidence interval can be computed by taking the reciprocals of the corresponding quantities for the risk difference.

Both Cole and Ozenne focused on estimating the ATE of the risk difference using CIFs. Both methods account for covariates affecting both treatment selection and censoring, with the possibility that the covariates affecting treatment selection may differ from those affecting censoring. Both methods estimate inverse probability weights separately for treatment and censoring and incorporate these weights into the estimators of the risk difference. The estimation of inverse probability of treatment weights is similar between the two approaches. However, the two approaches differ in estimation of the IPCWs. In Cole, the model for dropout (i.e., right censoring) is based on grouping individuals into discrete intervals of time based on the quintiles of censoring times and then using a pooled logistic regression model to estimate interval-specific IPCWs. It is worth noting that because Cole's estimator is based on the Aalen–Johansen estimator, it is valid even without the inclusion of censoring weights if censoring is noninformative, that is, independent of the covariates. In contrast, Ozenne's estimator is developed using a different approach that requires censoring weights even when censoring is noninformative. Furthermore, unlike Cole's discretization of time for estimating censoring weights, Ozenne's approach uses continuous time for estimating the censoring model. In the absence of the weights and without censoring, both approaches would reduce to the difference in the Aalen–Johansen estimator for the two treatment groups (without censoring) and would be identical.

The reporting of absolute measures of effect can be complemented by reporting relative measures of effect. The methods described above can be used to estimate the absolute risk of the outcome under treatment and control. Then, the corresponding relative risk can be computed as the ratio of these two risks. Either weighted Aalen–Johansen estimates of the CIF under treatment and control or Ozenne's estimates of the absolute risk under treatment and control can be used. As an appropriate variance estimator for the relative risk has not been described, one can use the bootstrap, as suggested by Cole and colleagues [15]. Alternatively, as Ozenne and colleagues developed an estimate of the standard error of the risk of the outcome under treatment and control, one can use the delta method to approximate the standard error for the relative risk when it is computed using Ozenne's estimators. Using the delta method, an estimate of the standard error of the log-relative risk would be $\sqrt{\left(\frac{\mathrm{se}(\hat{p}_0)}{\hat{p}_0}\right)^2 + \left(\frac{\mathrm{se}(\hat{p}_1)}{\hat{p}_1}\right)^2 - 2\frac{\mathrm{Cov}(p_0,p_1)}{\hat{p}_0\hat{p}_1}}$, where $\hat{p}_0$ and $\hat{p}_1$ denote the mean estimated risk under control and treatment, respectively, $\mathrm{se}(\hat{p}_0)$ and $\mathrm{se}(\hat{p}_1)$ denote the estimated standard errors of the estimated risk under control and treatment, respectively, while $\mathrm{Cov}(p_0,p_1)$ denotes the covariance between the estimated risk under control and treatment, respectively.

3.3 | Estimating the Relative Effect of Treatment on the Cause-Specific Hazard Function When Using IPTW

In the beginning of Section 3, we highlighted Dorn's dictum that the analysis of observational studies should reflect what would be done in a corresponding RCT. RCTs in clinical medicine with time-to-event outcomes, with or without competing risks, frequently report a hazard ratio, which denotes the relative change in the hazard of the outcome under treatment.

In the setting with time-to-event outcomes and competing risks, the cause-specific hazard function for the kth event type is a function of time (t) that is defined as $\lambda_k^{\mathrm{cs}}(t) = \lim_{\Delta t \to 0} \frac{\mathrm{Prob}(t \le T < t+\Delta t, D=k | T \ge t)}{\Delta t}$, where T is the time at which an event occurred and D is a variable denoting the type of event that occurred (with $D=k$ denoting that the kth event type has occurred). The cause-specific hazard function for the kth event type can be interpreted as the instantaneous rate of occurrence of the kth event type in subjects who are currently event-free (e.g., subjects for whom no event of any type has occurred).

A proportional cause-specific hazard model for the kth event type allows one to estimate the association of covariates with the cause-specific hazard function for the kth event type: $\lambda_k^{\mathrm{cs}}(t|\mathbf{X}) = \lambda_{0k}^{\mathrm{cs}}(t)\exp(\beta\mathbf{X})$, where $\lambda_{0k}^{\mathrm{cs}}(t)$ denotes the baseline cause-specific hazard function for the kth event type, and $\mathbf{X}$ denotes a vector of covariates. In practice, the cause-specific hazard model can be fit using standard statistical software for fitting the Cox proportional hazards model. To do so, one censors subjects who experience a competing event at the time the competing event occurs (e.g., a subject who experiences a competing event at time t_{ce} is treated as censored at time t_{ce} and is no longer under observation from that time onwards). The estimated regression coefficients can be interpreted as denoting the association of the covariate with the rate at which the event of interest occurs in subjects who are currently event-free.

When using IPTW using the propensity score, the relative effect of treatment on the hazard of the outcome of interest can be estimated using a cause-specific hazard model, in which the cause-specific hazard of the outcome of interest is regressed

on an indicator variable denoting treatment status. A weighted regression model is used that incorporates the inverse probability of treatment weights. As weighting has balanced the distribution of observed covariates between treatment groups, it is not necessary to adjust for other baseline covariates. Previous research has demonstrated that one should use either a robust variance estimator or the bootstrap, with the latter being preferable [4, 38]. The use of the weighted cause-specific hazard model allows investigators to test whether the rate of the occurrence of the outcome in subjects who are currently event-free is the same between treated and control subjects.

Two earlier articles examined the performance of IPTW using the propensity score to estimate marginal hazard ratios in the absence of competing risks [4, 38]. Given the equivalence between cause-specific hazard models and Cox proportional hazards model in which one censors on competing risks, we do not examine further the performance of IPTW to estimate marginal cause-specific hazard ratios in the current study.

4 | Case Study

We provide a case study to illustrate the application of IPTW using the propensity score in the presence of competing risks. The exposure of interest is statin prescribing at hospital discharge, while the outcome is death within 5 years of follow-up. Death was classified as due to cardiovascular causes or non-cardiovascular causes. The primary outcome of interest is cardiovascular death.

4.1 | Data Sources

We used data from the first phase of the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) Study, which collected detailed clinical data on patients hospitalized with acute myocardial infarction (AMI) between April 1, 1999 and March 31, 2001 at 103 hospitals in Ontario, Canada [39]. Data were obtained on patient demographics, vital signs, and physical examination at presentation, medical history, and results of laboratory tests. For the following analyses, we restricted the study sample to 10 063 patients who were discharged alive from the hospital.

Subjects were linked to the Vital Statistics database maintained by the Ontario Office of the Registrar General. This database contains information on the date of death and cause of death (based on ICD-9 codes) for residents of Ontario. Each subject was followed for 5 years from the date of hospital discharge for the occurrence of death. The primary outcome was death due to major cardiovascular disease (subsequently referred to as cardiovascular death) [40], whereas death due to other causes was treated as a competing risk (subsequently referred to as non-cardiovascular death). A total of 2966 (29.5%) patients died during the 5 years of follow-up. Of these, 1818 (61%) died of cardiovascular causes, while 1148 (39%) died of non-cardiovascular causes. No individuals were censored prior to 5 years, and all individuals who were event-free at 5 years were censored at that time.

The following nine predictor variables were used as baseline covariates: age, heart rate at hospital admission, systolic blood pressure at admission, initial serum creatinine, history of AMI, history of heart failure, ST-depression myocardial infarction, elevated cardiac enzymes, and in-hospital percutaneous coronary intervention (PCI). These variables were selected because they are components of the GRACE risk score for predicting mortality in patients with acute coronary syndromes [41]. The first four variables are continuous variables, while the last five are dichotomous risk factors. We standardized the four continuous variables so that they had a mean of zero and unit variance. The prevalences of the five binary variables were: history of AMI (22.5%), history of heart failure (4.1%), ST-depression myocardial infarction (48.0%), elevated cardiac enzymes (94.1%), and in-hospital PCI (1.1%).

We used discharge prescribing of a statin lipid-lowering agent as the exposure of interest. Of the 10 063 patients discharged alive from the hospital, 3359 (33.4%) received a prescription for a statin medication at hospital discharge.

4.2 | Statistical Analyses

We estimated the propensity score using a logistic regression model to regress statin prescribing at hospital discharge on the nine baseline covariates described above. The logistic regression model only included main effects for the nine covariates. Continuous variables were assumed to have a smooth nonlinear relationship with the log-odds of treatment, modeled using restricted cubic splines with five knots, using the knot locations suggested by Harrell [42]. From the fitted model we extracted the fitted propensity scores and computed inverse probability of treatment weights. We used weighted standardized differences to assess the balance in the nine baseline covariates between treated and control subjects in the weighted sample [43].

We used the methods described above to estimate absolute risk differences and relative risks for cardiovascular death due to statin prescribing at discharge. When using Cole's method, we did not incorporate censoring weights, since the only censoring that occurred was at 5 years (at the end of follow-up). When using doubly robust methods, the outcomes model was a cause-specific hazard model that included the nine baseline covariates and a binary variable denoting treatment status. As with the propensity score model above, the relationship between each of the continuous covariates and the cause-specific hazard of the outcome was modeled using restricted cubic splines with five knots. For both the IPTW-IPCW estimator and the AIPTW-IPCW estimator, the censoring distribution was modeled using a null Cox model (we also repeated the AIPTW-IPCW analysis using a Cox model that contained all the predictor variables in the treatment-selection model). The standard errors of the estimated risk differences were estimated using the variance estimator proposed by Ozenne and colleagues. For the IPTW-IPCW and AIPTW-IPCW estimators, we used the delta method to compute standard errors of the log-relative risk and constructed 95% confidence intervals using standard normal-theory methods. For the weighted Aalen–Johansen method, 95% confidence intervals for the time-specific relative risks were estimated using percentile bootstrap confidence intervals with 2000 bootstrap samples.

We fit a cause-specific hazard model in the weighted sample in which we regressed the cause-specific hazard of cardiovascular

death on an indicator variable denoting treatment status. We fit a marginal model with a robust variance estimator to account for within-subject homogeneity induced by nonuniform weighting [4].

We used the prodlim() function in the prodlim package (Version 2019.11.13) to obtain the weighted Aalen–Johansen estimate of the risk difference and the relative risk. We used the ate() function in the riskRegression package (Version 2020.12.08) for estimating risk differences using the IPTW-IPCW and AIPTW-IPCW estimators. The relative risks were computed using information extracted from the ate() function. The weighted cause-specific hazard model was estimated using the coxph function in the survival package (Version 3.2-11).

4.3 | Results of Empirical Analyses

Prior to weighting, the standardized differences for the nine covariates ranged from 0.03 to 0.37, with a median of 0.09. Unweighted standardized differences exceeded 0.10 for four of the baseline covariates (it has been suggested that standardized differences that are less than 0.10 denote negligible imbalance [44]). Thus, there is evidence of confounding, with the distribution of prognostically important baseline covariates differing between treated and control subjects. Excellent balance of baseline covariates between treated and control subjects was observed after the application of the inverse probability of treatment weights, with the absolute value of the standardized difference ranging from 0 to 0.02, with a median of 0.

The function for computing the AIPTW-IPCW estimates did not work. Accordingly, we only report the weighted Aalen–Johansen and IPTW-IPCW estimates of the risk difference and relative risk.

The estimated crude and weighted CIFs in treated and control subjects are shown in Figure 1. Post-discharge incidence of cardiovascular death was lower in treated subjects than in control subjects. After weighting, differences in the incidence of cardiovascular death were attenuated compared to the crude or unadjusted differences. The weighted Aalen–Johansen and IPTW-IPCW estimates of the CIF were essentially identical.

The crude differences in the risk of cardiovascular death due to discharge prescribing of statins were −0.046, −0.063, −0.073, −0.081, and −0.088 at 1, 2, 3, 4, and 5 years post-discharge, respectively. The weighted Aalen–Johansen estimates of the differences in the risk of cardiovascular death due to discharge prescribing of statins, along with associated 95% confidence intervals, were −0.021 (−0.033, −0.009), −0.028 (−0.041, −0.013), −0.033 (−0.047, −0.017), −0.037 (−0.052, −0.020), and −0.035 (−0.052, −0.019) at 1, 2, 3, 4, and 5 years post-discharge, respectively. The NNTs to avoid one death due to cardiovascular causes at 1, 2, 3, 4, and 5 years post-discharge were 48, 36, 30, 27, and 29, respectively. The IPTW-IPCW estimates of the differences in the risk of cardiovascular death due to discharge prescribing of statins,

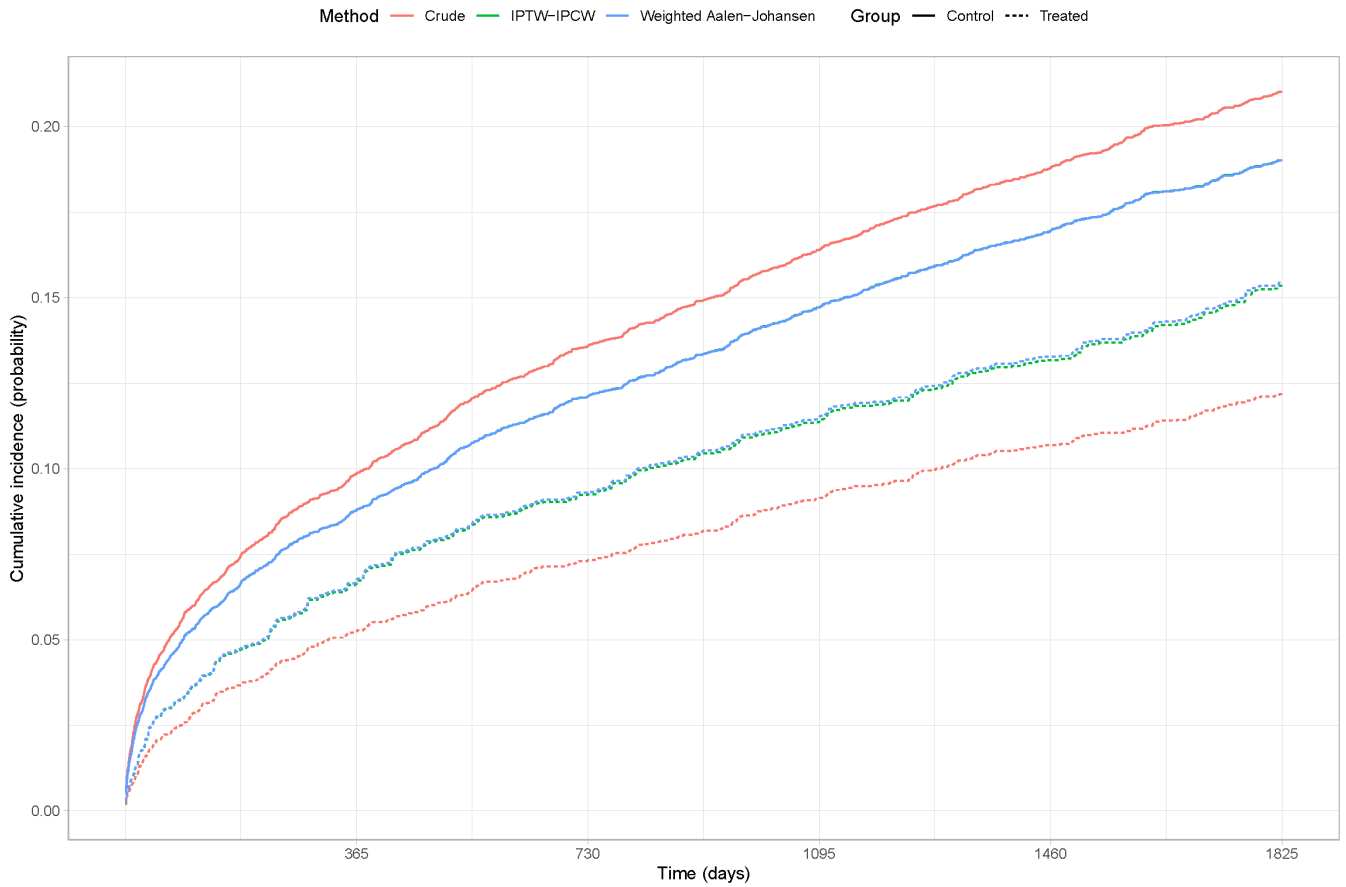

FIGURE 1 | Cumulative incidence of cardiovascular death in treated and control subjects: Case study.

along with associated 95% confidence intervals, were −0.022 (−0.034, −0.010), −0.029 (−0.042, −0.015), −0.034 (−0.048, −0.019), −0.038 (−0.053, −0.022), and −0.037 (−0.053, −0.020) at 1, 2, 3, 4, and 5 years post-discharge, respectively. The NNTs to avoid one death due to cardiovascular causes at 1, 2, 3, 4, and 5 years post-discharge were 45, 34, 29, 26, and 27, respectively.

The crude relative risks were 0.531, 0.540, 0.556, 0.569, and 0.579 at 1 through 5 years post-discharge respectively. The weighted Aalen–Johansen estimates of the relative risks were 0.759 (0.634, 0.898), 0.770 (0.669, 0.887), 0.778 (0.688, 0.880), 0.783 (0.700, 0.876), and 0.814 (0.734, 0.901) at 1 through 5 years, respectively. The IPTW-IPCW estimates of the relative risks were 0.753 (0.634, 0.894), 0.764 (0.664, 0.880), 0.772 (0.682, 0.873), 0.777 (0.695, 0.868), and 0.807 (0.729, 0.894) at 1 through 5 years post-discharge, respectively. Recall that 95% confidence intervals for the weighted Aalen–Johansen estimates of the relative risk were constructed using a bootstrap procedure while those for the IPTW-IPCW estimates were constructed using the delta method. When we used a bootstrap procedure for the IPTW-IPCW estimate of relative risk, we obtained very similar confidence intervals to those constructed using the delta method (the largest absolute difference between the endpoints of the eight 95% confidence intervals was 0.004).

The estimated cause-specific hazard ratio for cardiovascular death was 0.779 (95% confidence interval: 0.689, 0.880) (when using the bootstrap with 2,000 bootstrap replicates, the corresponding percentile-based 95% bootstrap confidence interval was (0.698, 0.870)). Thus, statin prescribing at discharge decreased the rate of cardiovascular death in subjects who were currently alive by 22.1%.

Our reporting of the effect of statin prescribing at hospital discharge on the incidence of cardiovascular death mirrored what one would expect in the report of a comparable RCT: (i) the reporting of the cumulative incidence of the outcome over time in treated and control subjects; (ii) the absolute risk reduction due to treatment at specific durations of time; (iii) the reporting of NNTs at specific durations of time; and (iv) the effect of treatment on the instantaneous cause-specific hazard of the outcome.

5 | Monte Carlo Simulations

In this section, we describe a set of complex Monte Carlo simulations that were designed to assess the performance of different methods for estimating risk differences and relative risks at specific durations of time in settings with competing risks.

The design of our simulations was informed by analyses conducted on the data described in the case study above. In Section 5.1, we describe the data and statistical analyses that were used to estimate parameters for the subsequent data-generating process. In Section 5.2, we describe the data-generating process that was used to simulate survival data from a specified subdistribution hazard model. In Section 5.3, we describe the statistical analyses that were conducted on the simulated data. In Section 5.4, we explain how the results of the analyses were summarized across simulation replicates. Finally, in Section 5.5, we describe the factors that were allowed to vary in the Monte Carlo simulations. These simulations are similar in design to previous simulations that we used to examine the effect of the number of events per variable (EPVs) on the accuracy of estimation of the regression coefficients if Fine–Gray subdistribution hazard models and on the use of propensity score matching with competing risks [14, 45].

5.1 | Data Sources and Empirical Statistical Analyses

The design of the Monte Carlo simulations was informed by analyses conducted on the data which were used in the case study and which were described in Section 4.1. We used logistic regression to regress statin prescribing at hospital discharge on the nine covariates described above. The vector of regression coefficients for this logistic regression model is denoted by $\boldsymbol{\alpha}$. These estimated regression coefficients will be used in the treatment-selection model in our data-generating process.

We used a Fine–Gray subdistribution hazard regression model to regress the subdistribution hazard of cardiovascular death on the nine baseline covariates described above and an indicator variable denoting statin prescribing at hospital discharge. The vector of regression coefficients for the Fine–Gray subdistribution hazard model for cardiovascular death is denoted by $\boldsymbol{\beta}$. We fitted a second Fine–Gray subdistribution hazard model to regress the subdistribution hazard of non-cardiovascular death on the nine baseline covariates described above. The vector of regression coefficients from this model is denoted by $\boldsymbol{\gamma}$. These two vectors of regression coefficients will be used in generating outcomes in our data-generating process.

5.2 | Data-Generation Process

We simulated data for a large super-population of 1 000 000 subjects. Using a data-generating process whose parameters were obtained by analyses conducted in the EFFECT sample resulted in the structure of this super-population reflecting that of the EFFECT sample.

5.2.1 | Simulation of Baseline Covariates

For each subject in the super-population, we simulated nine baseline covariates, such that the distribution of the baseline covariates would be similar to that of the nine baseline covariates described above. Four of the simulated covariates were continuous and were drawn from independent standard normal distributions (since the four continuous covariates had been standardized to have a mean of zero and unit variance in the empirical analyses described above). Five of the simulated covariates were binary and were drawn from independent Bernoulli distributions with parameters equal to the five prevalences described in Section 5.1. Let $X_1, \ldots, X_9$ denote the nine simulated baseline covariates, where the first four are continuous and the last five are binary.

5.2.2 | Simulation of Treatment Status

We simulated a treatment status (Z_i) for each subject from a Bernoulli distribution: $Z_i \sim \text{Be}(p_{\text{treat}})$, where $\text{logit}(p_{\text{treat}}) =$

$\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_9 X_9$. The nine regression coefficients $(\alpha_1, \ldots, \alpha_9)$ for the nine baseline covariates in the treatment-selection model were equal to the regression coefficients estimated in the empirical analyses described above. A bisection approach was used to determine the intercept for the treatment-selection model (α_0) so as to induce a desired prevalence of treatment in the large super-population (see below for the target prevalences of treatment) [46]. We thus simulated a treatment status for each subject such that the relationship between the nine baseline covariates and the odds of treatment reflected what was observed in the EFFECT data. The only difference was that we modified the prevalence of treatment in the simulated data. This allowed us to examine the effect of the prevalence of treatment on the performance of IPTW using the propensity score.

5.2.3 | Generation of the Event Type That Occurred (Type 1 vs. Type 2 Event) and the Event Time

Let the parameter p denote the proportion of subjects with covariates equal to zero (i.e., the continuous variables are equal to zero and binary covariates are set to the reference level) who experience the event of interest as $t \to \infty$. We generated event types using a method described in detail previously [45]. In generating event types, we used the vector $\boldsymbol{\beta}$, obtained in Section 5.1, which is equal to the effect of the nine covariates and treatment status on the incidence of cardiovascular death. We allowed p to take on a range of plausible values (see the section below describing the design of the Monte Carlo simulations).

We simulated time-to-event outcomes conditional on the simulated failure type using a data-generating process described in detail elsewhere [45]. This data-generating process is based on a method of indirect simulation described by Beyersmann et al. [47] (Section 5.3), which in turn is based on an approach described by Fine and Gray [48]. This approach ensures that the sum of the two CIFs equals exactly 1 as t goes to infinity. In doing so, one needs only to specify the underlying subdistribution hazard function for the event of interest and the hazard function for the conditional distribution of the event time given failure from a competing event, rather than the cause-specific hazard functions for the two event types. In simulating event times, we used the two vectors of regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ that were estimated in the empirical analyses above, where the parameter $\boldsymbol{\beta}$ was used for the subdistribution for the event of interest and the parameter $\boldsymbol{\gamma}$ was used for the conditional distribution for the competing event. However, the vector $\boldsymbol{\beta}$ was modified so that the regression coefficient for treatment was replaced with the logarithm of the desired conditional subdistribution hazard ratio. Thus, we were simulating data with a specified conditional subdistribution hazard ratio for treatment (see below for the different values that this conditional subdistribution hazard ratio could take). For each subject in the super-population, two potential outcomes were simulated: the potential outcome under control and the potential outcome under treatment [18] (this will allow calculation of the true risk differences and relative risks at specified durations of time — see the subsequent section for a description of how this was accomplished). Each subject's observed event time was set to be equal to the potential outcome for the treatment that the subject actually received (i.e., if a subject was treated, then the observed event time was set equal to the potential outcome under treatment, while if a subject was a control subject, then the observed event time was set equal to the potential outcome under control). Note that this data-generating process ensures that an event time will be generated for each subject and that the event type will be either the primary event or the competing event. We then generated a random censoring time for each subject from an exponential distribution. For each simulated subject, the observed survival time was the minimum of the event time and the censoring time. We used a bisection approach to determine the rate parameter for the exponential distribution so that 20% of subjects were censored.

5.3 | Statistical Analyses in Simulated Dataset

5.3.1 | Determination of the True Risk Differences and Relative Risks in the Super-Population

The process for generating time-to-event outcomes used a model conditional on both treatment and other covariates. We computed the true marginal risk differences and relative risks at specified times in the super-population. We determined the 20th, 40th, 60th, and 80th percentiles of observed event times (i.e., time to either the primary event type or the time to the competing risk) for subjects in the super-population. We refer to these four times as T_{20}, T_{40}, T_{60}, and T_{80}. These will be the four times at which we will estimate risk differences and relative risks.

For each of the 1 000 000 subjects in the super-population, we used the simulated potential outcome under control and obtained the Aalen–Johansen estimate of the CIF. Using the estimated CIF, we obtained the estimate of the risk of the primary outcome at T_{20}, T_{40}, T_{60}, and T_{80}. We denote these four estimates of risk as $R(0,T_{20})$, $R(0,T_{40})$, $R(0,T_{60})$, and $R(0,T_{80})$. We then repeated the process using the 1 000 000 values of the simulated potential outcome under treatment to obtain $R(1,T_{20})$, $R(1,T_{40})$, $R(1,T_{60})$, and $R(1,T_{80})$. The true values of the risk difference at the four times are defined as $R(1,T_j) - R(0,T_j)$, for $j = 20, 40, 60$, and 80. Similarly, the true values of the relative risk at the four times are defined as $R(1,T_j)\,/(0,T_j)$.

5.3.2 | Estimation of Treatment Effects Using Inverse Probability of Treatment Weighting

From the super-population we drew a random sample of size N without replacement (see below for the values of N that were used). In this random sample, we estimated the propensity score using logistic regression to regress the binary treatment variable on the nine baseline covariates. We then computed the inverse probability of treatment weights.

We estimated risk differences and relative risks at the four time points determined above. We compared three methods of estimating the risk difference and relative risk: (1) using a weighted Aalen–Johansen estimate of the CIF under treatment and control (i.e., the method proposed by Cole and colleagues); (2) using the IPTW-IPCW estimator proposed by Ozenne and colleagues [17]; and (3) using the AIPTW-IPCW estimator proposed by Ozenne and colleagues. When using the weighted Aalen–Johansen

method, we did not incorporate IPCWs since the simulation process induced random censoring. As noted above, when censoring is random, the IPCWs can be excluded. When using the doubly robust estimator, the outcomes model was a Cox model in which the hazard of the outcome was regressed on the nine simulated baseline covariates and the binary treatment variable. For both the IPTW-IPCW and AIPTW-IPCW estimators, the censoring distribution was modeled using a null Cox model. When estimating the risk difference using IPTW-IPCW and AIPTW-IPCW, we used the variance estimators proposed by Ozenne and colleagues. When using the weighted Aalen–Johansen estimator for the risk difference or the relative risk, we obtained only the point estimate and did not obtain an estimate of its standard error. The variance estimator proposed by Ozenne and colleagues is very computationally intensive. Thus, we only used this variance estimator when the sample size was 1000 (see below for values of the sample size that were used).

For those methods and scenarios in which we estimated standard errors, we also computed 95% confidence intervals for the risk difference using standard normal-theory methods in each of the random samples.

This process was repeated 1000 times for each scenario. Thus, 1000 random samples of size N were drawn from the super-population, and these statistical analyses were conducted in each of the 1000 random samples.

5.4 | Summarizing the Results of the Simulations

The risk difference and the relative risk are the target estimands. The relative bias of the estimated risk difference and relative risk was defined as: $100 \times \frac{\frac{1}{1000}\sum_{i=1}^{1000}\widehat{\phi}_i - \phi}{\phi}$, where ϕ denotes the true value of the estimand determined during the data-generating process, and $\widehat{\phi}_i$ denotes the value of the estimate in the ith random sample.

The empirical standard error of the estimated treatment effect was computed as the standard deviation of the estimated treatment effect across the 1000 simulation replicates. We complement the empirical standard error by reporting the relative percent increase in precision compared to a reference method. For a given estimation method, denoted by "A," the empirical standard error was estimated as the standard deviation of the estimated treatment effect across the 1000 simulation replicates: $\text{EmpSE}_\text{A} = \text{SD}(\widehat{\phi})$. For each estimator, we computed the relative percent increase in precision compared to the AIPTW-IPCW method. This quantity was defined as: $100\left(\left(\frac{\text{EmpSE}_\text{AIPTW-IPCW}}{\text{EmpSE}_\text{A}}\right)^2 - 1\right)$, where "A" denotes the method being compared to AIPTW-IPCW [49]. If this quantity is less than 0, then the AIPTW-IPCW method had a smaller empirical standard error than method "A." If this quantity is greater than 0, then the AIPTW-IPCW method had a larger empirical standard error than did method "A.". Everything else being equal, one would prefer a method with a smaller empirical standard error. When estimating relative risks, we determined the empirical standard error of the log-relative risk rather than of the relative risk itself.

In those scenarios and for those methods in which standard errors were estimated, the relative percent error in the estimated standard error of the estimated risk difference (or log-relative risk) was computed as: $100 \times \left(\frac{\frac{1}{1000}\sum_{i=1}^{1000}\text{se}\left(\widehat{\phi}_i\right)}{\text{SD}(\widehat{\phi})} - 1\right)$, where $\text{se}\left(\widehat{\phi}_i\right)$ denotes the estimated standard error in the ith random sample and $\text{SD}(\widehat{\phi})$ denotes the standard deviation of the estimated risk difference (or log-relative risk) across the 1000 random samples [49]. If the relative error is equal to zero, then the estimated standard error is correctly estimating the standard deviation of the sampling distribution of the estimated treatment effect. If the relative error is less than zero, then the estimated standard errors are underestimating the standard deviation of the sampling distribution of the estimated treatment effect. If the relative error is greater than zero, then the estimated standard errors are overestimating the standard deviation of the sampling distribution of the estimated treatment effect. Empirical coverage rates of estimated 95% confidence intervals were computed as the proportion of estimated confidence intervals that contained the true value of the treatment effect that was specified in the data-generating process.

5.5 | Design of the Monte Carlo Simulations

Our Monte Carlo simulations employed a full factorial design in which four factors were allowed to vary: (i) the size of the random samples from the super-population; (ii) the prevalence of treatment in the super-population; (iii) p (the proportion of subjects with covariates equal to zero who experience the primary event of interest as $t \to \infty$); and (iv) the true conditional subdistribution hazard ratio for treatment. The size of the random samples took three values: 1000, 2000, and 5000. The prevalence of treatment took five values: 0.1, 0.2, 0.3, 0.4, and 0.5. The parameter p was allowed to take three values: 0.25, 0.50, and 0.75. In doing so, we examined a range of plausible values for p, including a scenario in which the primary event was experienced by a higher proportion of subjects than the competing event, and a scenario in which the converse was true. The true subdistribution hazard ratio took three values: 1, 2, and 3. Thus, we examined 135 ($3 \times 5 \times 3 \times 3$) different scenarios.

The simulations were conducted using the R statistical software package [50] (Version 3.6.3). The prodlim() function in the prodlim package (Version 2019.11.13) was used to obtain the weighted Aalen–Johansen estimate of the risk difference and the relative risk. The ate() function in the riskRegression package (Version 2020.12.08) was used for estimating risk differences and relative risks using the IPTW-IPCW and AIPTW-IPCW estimators.

6 | Monte Carlo Simulations:Results

We present results separately for estimation of risk differences and estimation of relative risks.

6.1 | Estimation of Risk Differences

6.1.1 | Relative Bias of Estimated Risk Differences

The distributions of the relative bias across the 135 scenarios are described in the top row of panels in Figure 2. There is one

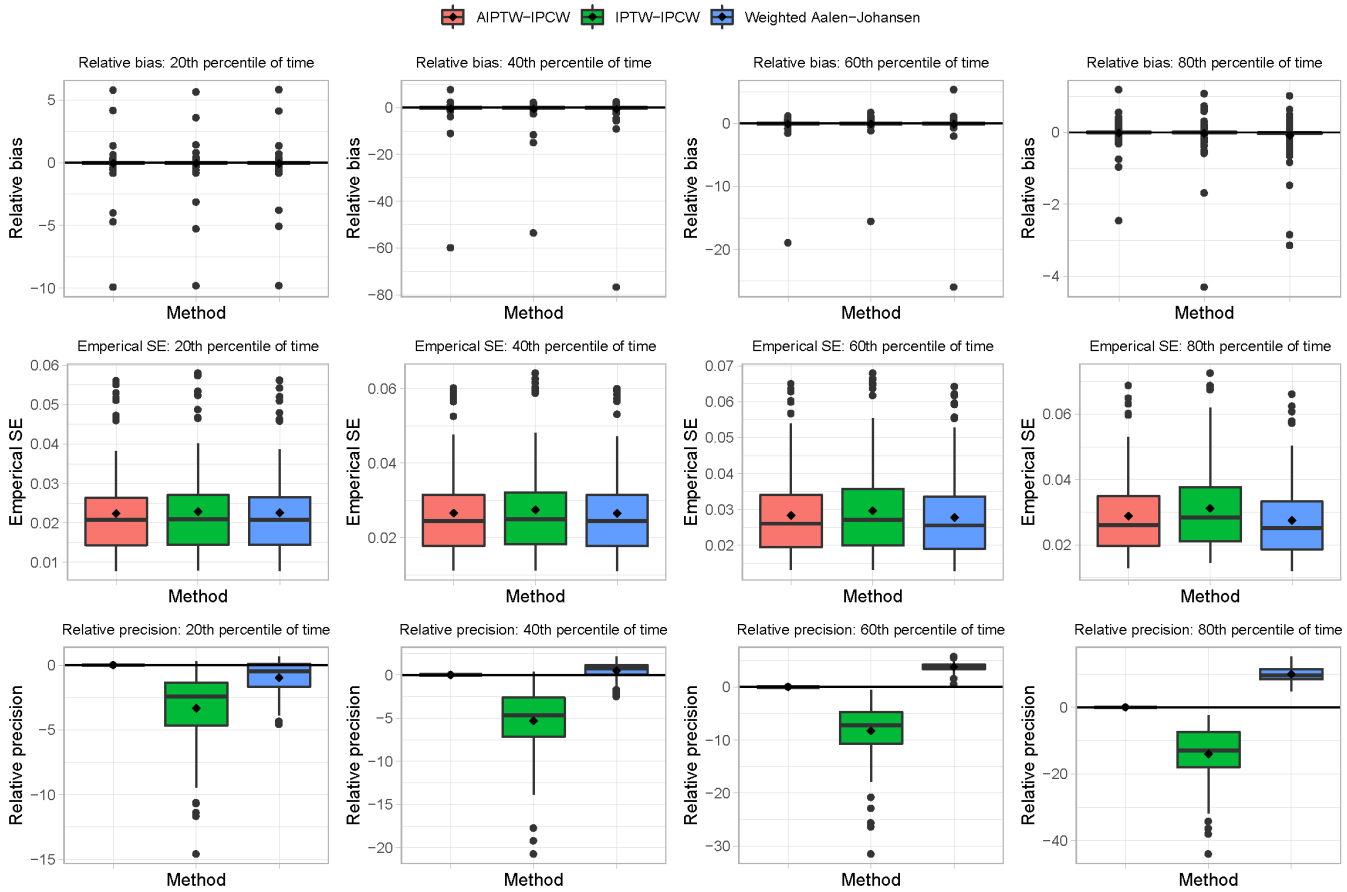

FIGURE 2 | Distribution of performance metrics across simulation scenarios: Risk differences.

panel for each of the four times at which risk differences were estimated. Each panel displays side-by-side boxplots describing the distribution of relative bias for each of the three estimation methods across the 135 scenarios. On average, each of the three methods resulted in unbiased estimation of the true risk difference.

Side-by-side boxplots allow a comparison of the distribution of a performance metric between different estimation methods. However, they do not allow one to compare the relative performance of the different methods in a specific scenario. To enable such a comparison, Figure 3 presents nested loop plots, which compare the relative bias between methods within each of the 135 scenarios [51]. The nested loop plot has one loop for each of the four factors in the design of the simulations (N: sample size; sd.hr.: subdistribution hazard ratio; p: proportion of events that are the primary event of interest; prop. treat: the prevalence of treatment [see the description of the four factors in the design of the Monte Carlo simulations]). The four factors that varied in the simulations are denoted by the top four lines in each panel. The top line, representing the outer loop, is a step function with three steps, denoting the three values that sample size (N) can take: 1000, 2000, and 5000. The text above this step function identifies the value of the factor associated with each step. Thus, the first step denotes a sample size of 1000. The second line (i.e., the line below the step function for sample size) is a step function that repeats three times and represents the values of the subdistribution hazard ratio for treatment (sd.hr). The number of times that the step function repeats is equal to the number of levels of the previous factor (sample size). The first step in each step function denotes $HR_{sd} = 1$, the second step in each step function denotes $HR_{sd} = 2$, while the third step denotes $HR_{sd} = 3$. The third line (i.e., the line below the function for sd.hr) is a step function that repeats nine times (the number of combinations of the levels for the two step functions above $= 3 \times 3$). The fourth line (i.e., the line below the step function for the third factor) is a step function that repeats 27 times (the number of combinations of the levels for the three step functions above $= 3 \times 3 \times 3$). One can draw 135 vertical lines on the figure, with each vertical line intersecting the four step functions in a unique way (e.g., there is one vertical line that intersects the first step function at $N = 1000$, the second step function at sd.hr. $= 1$, the third step function at $p = 0.25$, and the fourth step function at prop. treat $= 0.1$; this vertical line would be the left-most of the 135 vertical lines). Below these four step functions are three lines denoting the value of relative bias for the three estimation methods when the levels of the four factors are as denoted by the four step functions directly above. The 135 vertical lines described above would intersect each of these three lines. The height at which they intersect each of these lines denotes the value of the performance metric for the scenario identified by the top four step functions. In general, all three methods resulted in unbiased estimation of the true risk difference. Furthermore, differences between the methods tended to be minimal.

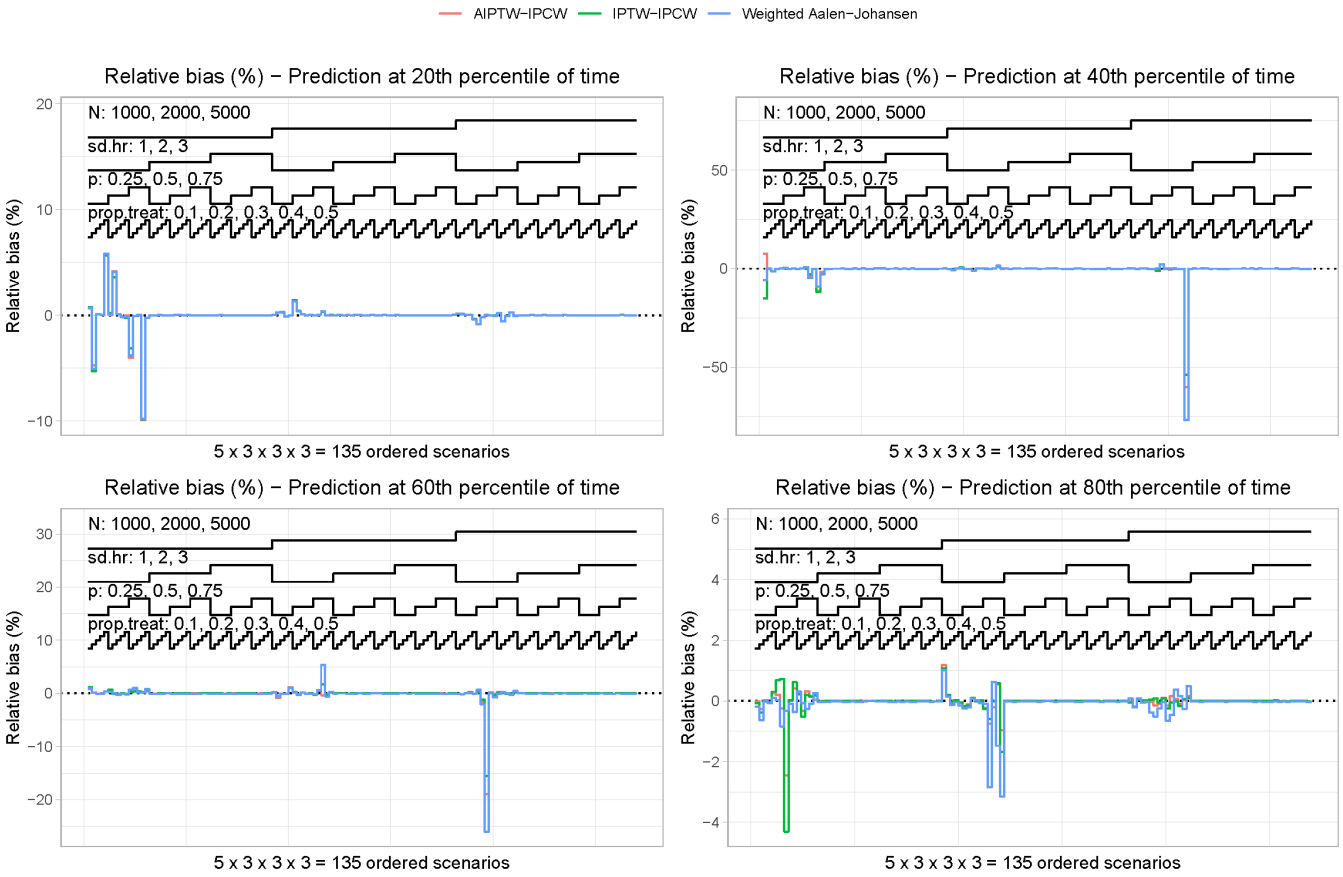

FIGURE 3 | Relative bias in estimated risk differences.

6.1.2 | Empirical Standard Errors and Estimation of Standard Errors of Estimated Risk Differences

The distribution of the empirical standard errors for each method across the 135 scenarios is reported in the middle row of panels in Figure 2. On average, IPTW-IPCW tended to result in estimates with slightly larger empirical standard errors compared to the other two methods. The empirical standard error is compared between methods across the 135 scenarios in Figure 4 using nested loop plots. Across many scenarios, the IPTW-IPCW method tended to result in estimates with marginally larger empirical standard errors compared to the other two methods.

The relative precision of the weighted Aalen–Johansen method and the IPTW-IPCW method compared to AIPTW-IPCW is reported in the lower row of panels in Figure 2 and using nested loop plots in Figure 5. These analyses confirm those in the preceding paragraph. When making predictions at the 20th percentile of event time, AIPTW-IPCW tended to result in estimates with the smallest empirical standard error. When making predictions at the 40th, 60th, and 80th percentiles of event time, the weighted Aalen–Johansen estimator tended to result in estimates with the smallest empirical standard error.

Standard errors were estimated using the IPTW-IPCW and AIPTW-IPCW methods, and, due to computational reasons, only in those scenarios in which the sample size was 1000. The distribution of the relative percent error in the estimated standard errors for each of these two methods and at each of the four times at which predictions were made is described in the four side-by-side boxplots in the upper row of panels in Figure 6. While differences between the two methods were minor, IPTW-IPCW tended to result in estimates with slightly less error. In general, both methods resulted in very minor underestimation of the standard deviation of the sampling distribution of the estimated relative risk. Nested loop plots comparing the relative percent error in the estimated standard errors are reported in Figure 7. Differences between the two methods were negligible.

6.1.3 | Coverage of 95% Confidence Intervals for Risk Differences

Empirical coverage rates of 95% confidence intervals were only assessed for the IPTW-IPCW and AIPTW-IPCW methods and for those scenarios in which the sample size was 1000. The distribution of empirical coverage rates across the simulation scenarios is reported using side-by-side boxplots in the lower row of panels in Figure 6. Due to our use of 1000 simulation replicates, any empirical coverage rate that is less than 0.9365 or greater than 0.9635 would be statistically significantly different from the advertised rate of 0.95, based on a standard normal theory test. We have added horizontal lines denoting these thresholds to the panels. In general, both methods tended to result in estimated confidence intervals whose empirical coverage rates did not differ from the advertised rate. When estimating risk differences at the 80th percentile of event times, IPTW-IPCW tended to have marginally better performance than AIPTW-IPCW. Nested loop

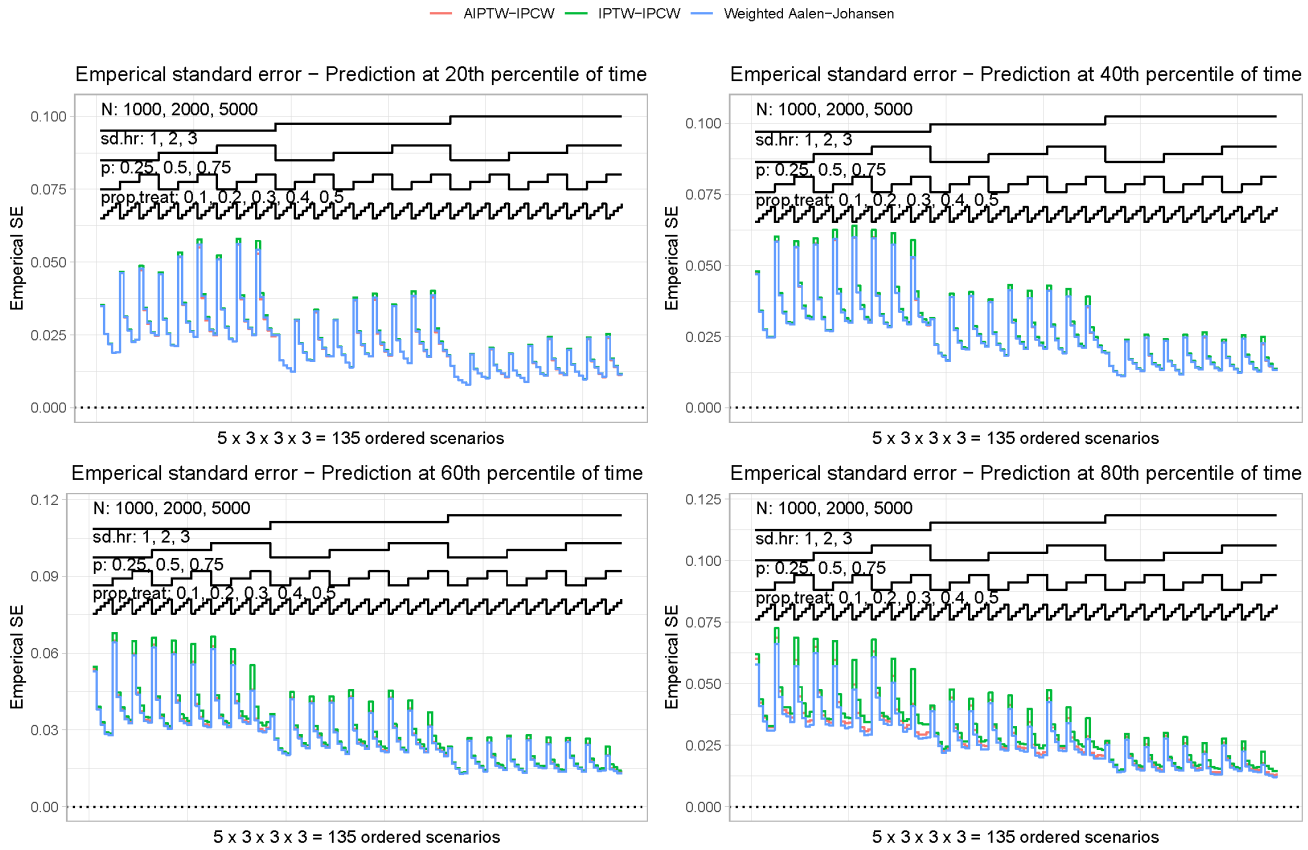

FIGURE 4 | Empirical standard error of estimated risk differences.

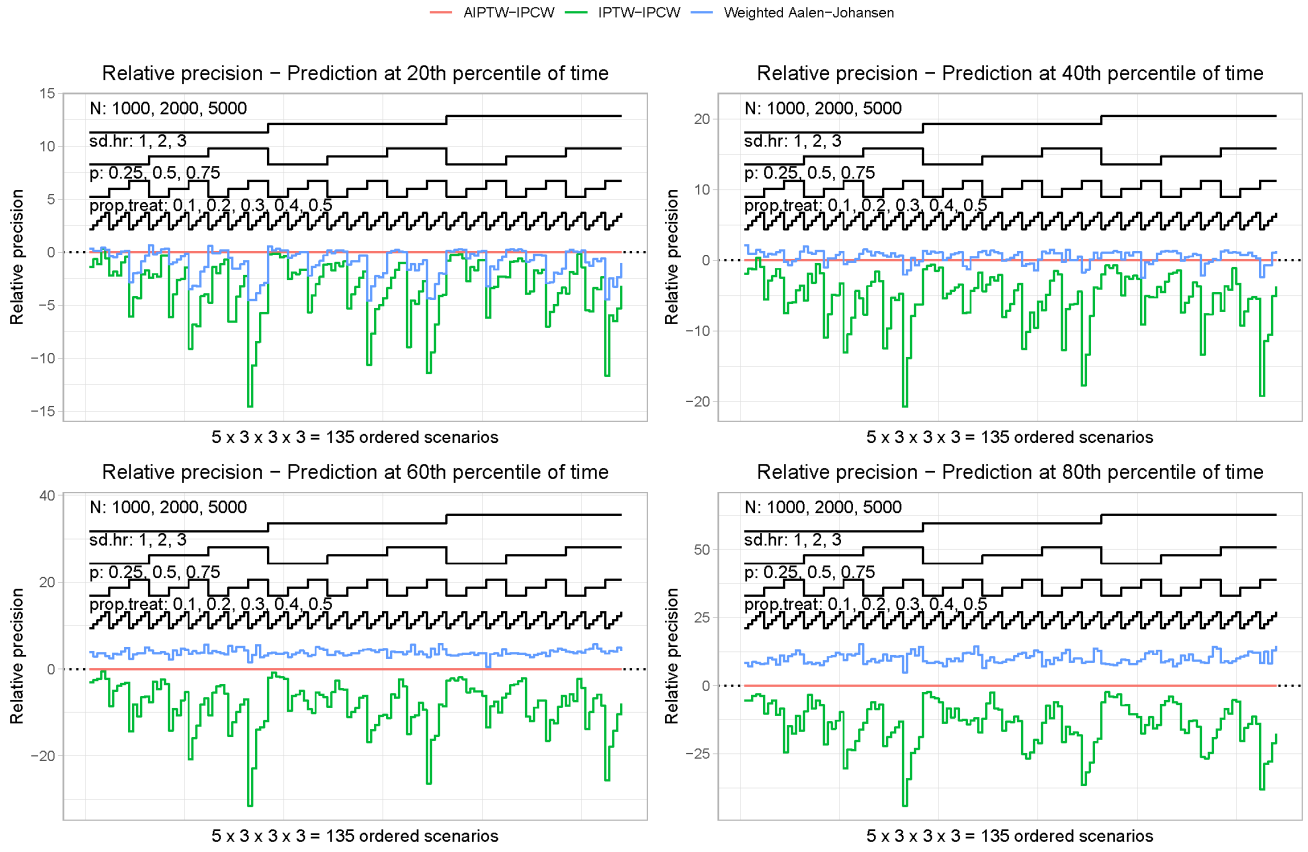

FIGURE 5 | Relative precision of estimated risk differences.

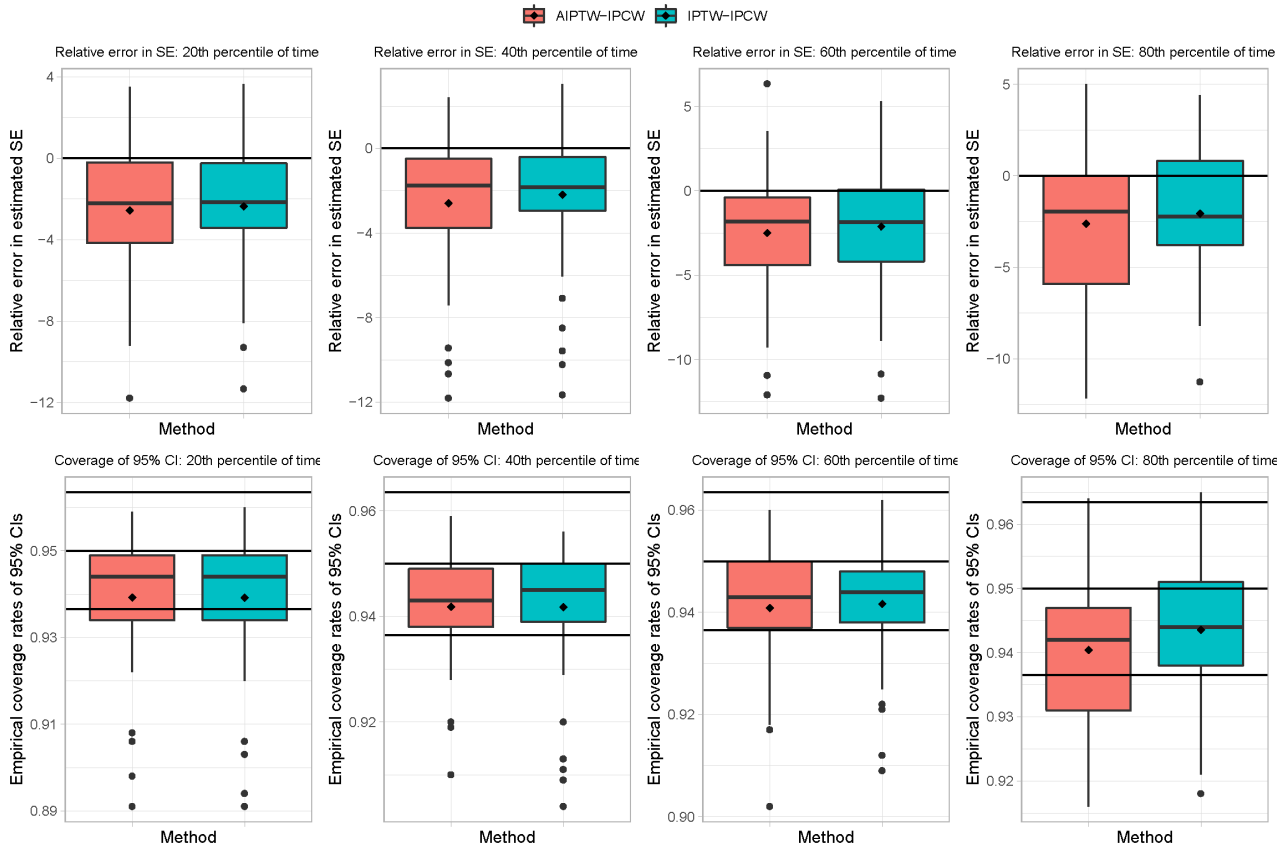

FIGURE 6 | Distribution of performance metrics for estimation of standard errors of risk differences.

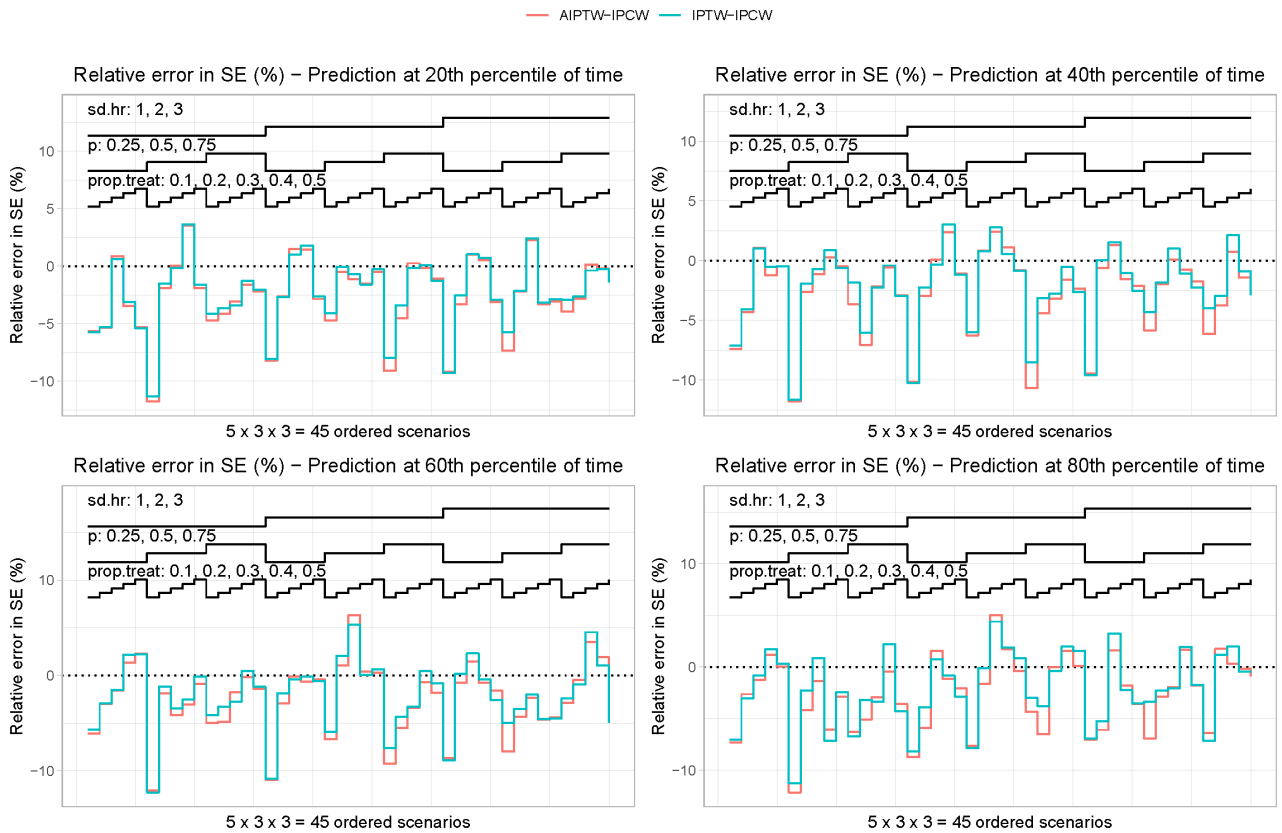

FIGURE 7 | Relative error in estimated standard error of the estimated risk difference.

plots comparing the performance of the two methods at each simulation scenario are reported in Figure 8.

6.2 | Estimation of Relative Risks

6.2.1 | Relative Bias of Estimated Relative Risks

The relative bias of estimated risk differences is reported in the top row of panels in Figure 9 (side-by-side boxplots) and in Figure 10 (nested loop plots). All three methods resulted in essentially unbiased estimation of the true relative risk.

6.2.2 | Empirical Standard Errors and Estimation of Standard Errors of Estimated Log-Relative Risks

The empirical standard errors of the estimated log-relative risks are reported in the middle row of panels in Figure 9 (side-by-side boxplots) and in Figure 11 (nested loop plots). The relative precision of each method compared to the AIPTW-IPCW method is reported in the bottom row of Figure 9 (side-by-side boxplots) and Figure 12 (nested loop plots). The IPTW-IPCW method tended to result in slightly larger empirical standard errors compared to the other two methods. The relative performance of the AIPTW-IPCW method and the weighted Aalen–Johansen method varied according to the time at which relative risk was estimated. However, on average, differences between these two methods were minimal.

Standard errors of the log-relative risk were estimated using the IPTW-IPCW and AIPTW-IPCW methods and, due to computational reasons, only in those scenarios in which the sample size was 1000. The distribution of the relative percent error in the estimated standard errors for each of these two methods and at each of the four times at which predictions were made is described in the four side-by-side boxplots in the upper row of panels in Figure 13. Although differences between the two methods were minor, IPTW-IPCW tended to result in estimated standard errors with slightly less error. In general, both methods resulted in minor underestimation of the standard deviation of the sampling distribution of the estimated relative risk. Nested loop plots comparing the relative percent error in the estimated standard errors are reported in Figure 14. Differences between the two methods were negligible.

6.2.3 | Coverage of 95% Confidence Intervals for Relative Risks

Empirical coverage rates of 95% confidence intervals were only assessed for the IPTW-IPCW and AIPTW-IPCW methods and for those scenarios in which the sample size was 1000. The distribution of empirical coverage rates across the simulation scenarios is reported using side-by-side boxplots in the lower row of panels in Figure 13. Due to our use of 1000 simulation replicates, any empirical coverage rate that is less than 0.9365 or greater than 0.9635 would be statistically significantly different from the advertised rate of 0.95 using a standard normal theory test. We

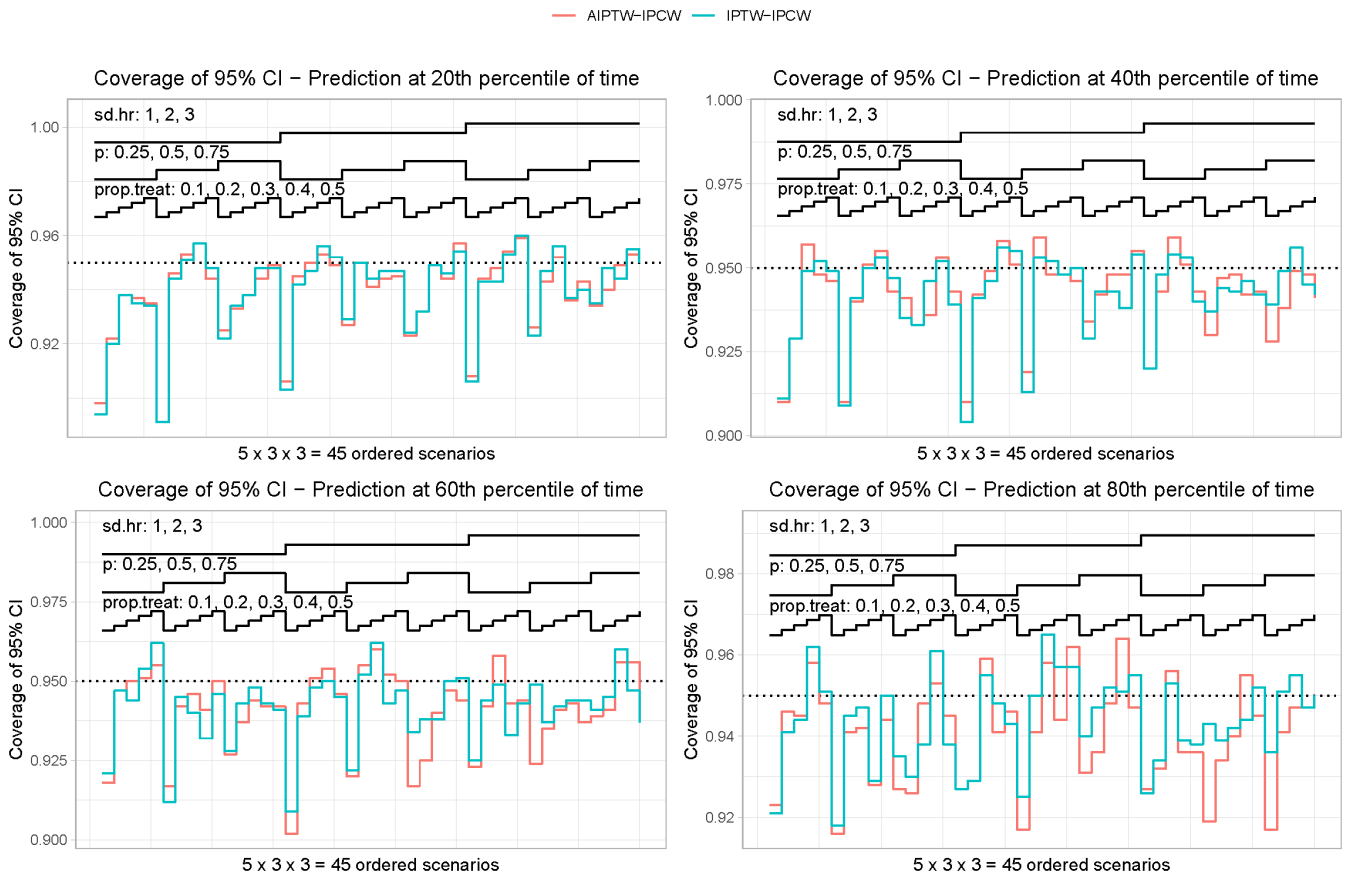

FIGURE 8 | Empirical coverage rates of 95% confidence intervals for estimated risk differences.

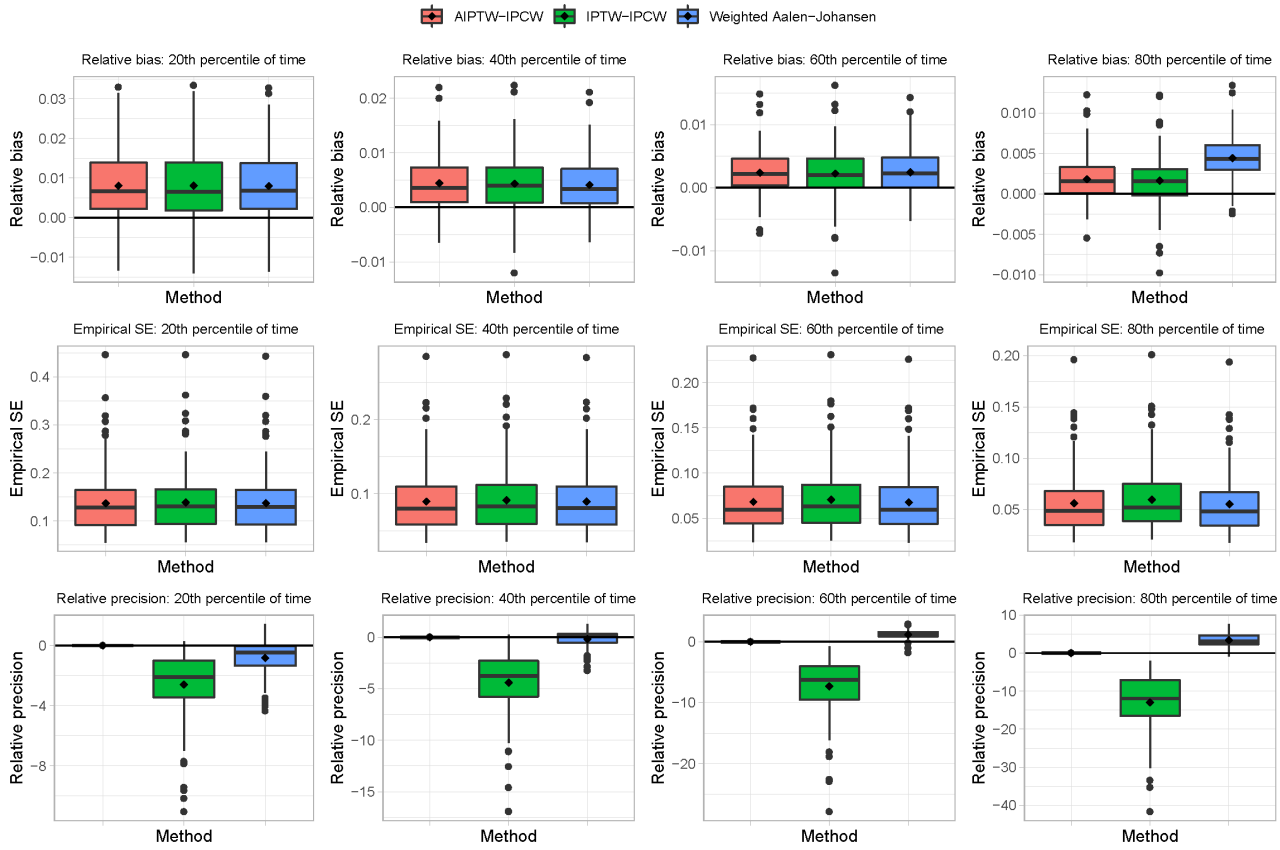

FIGURE 9 | Distribution of performance metrics across simulation scenarios: Relative risks.

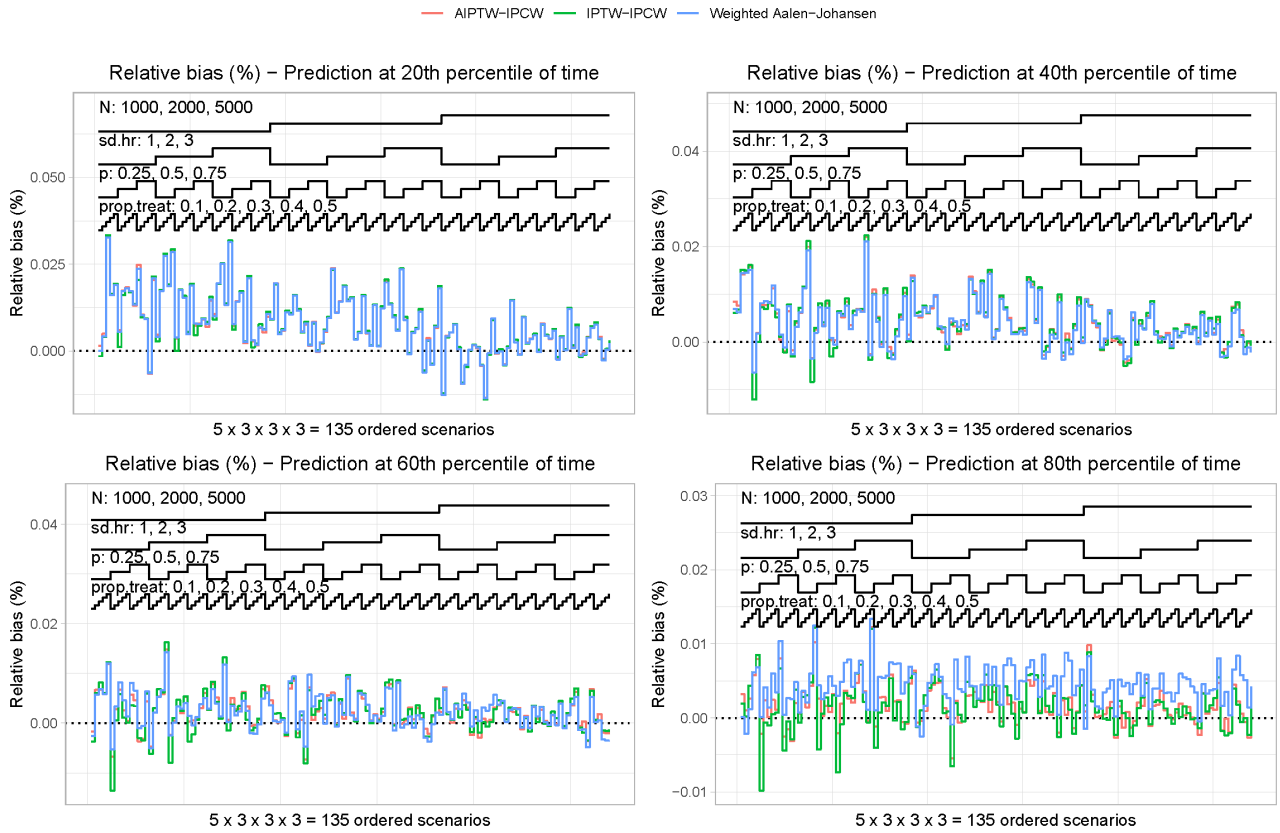

FIGURE 10 | Relative bias in estimated relative risk.

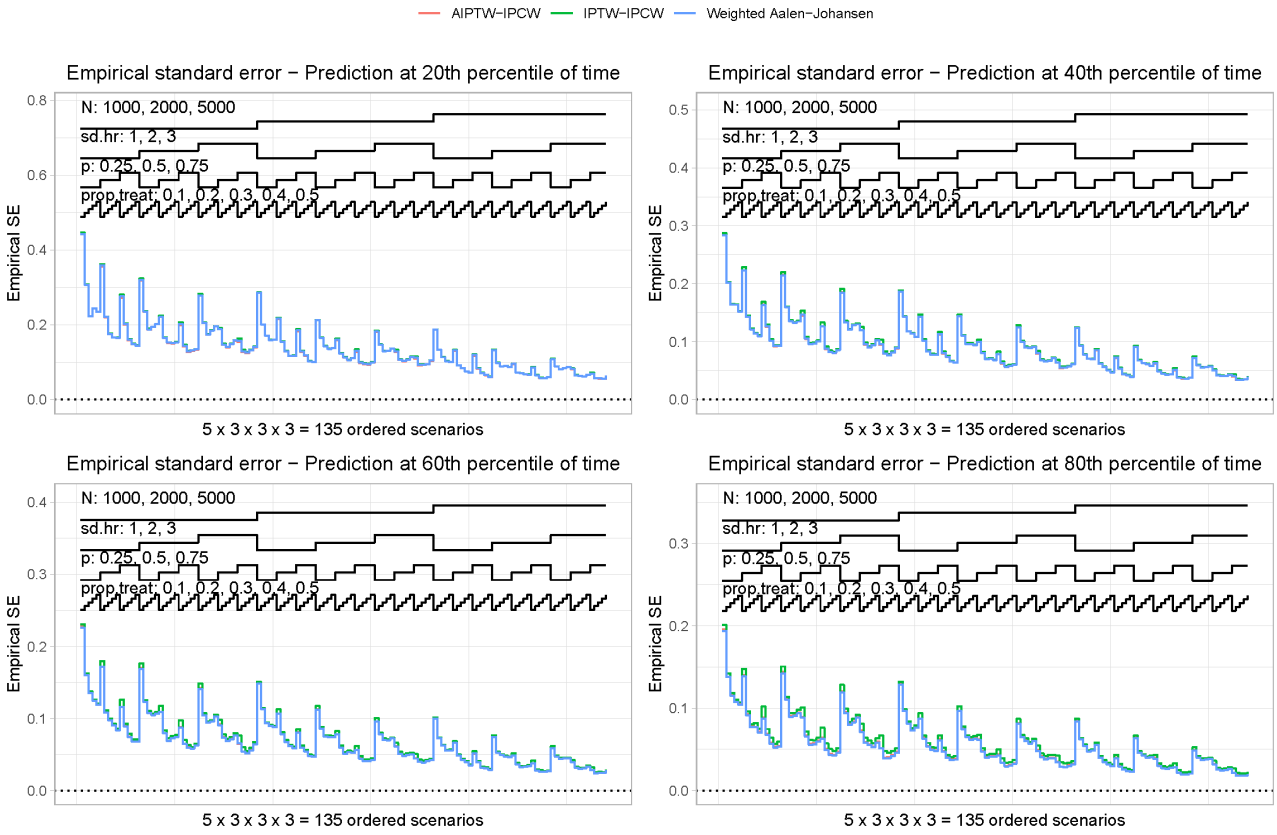

FIGURE 11 | Empirical standard error of estimated log-relative risk.

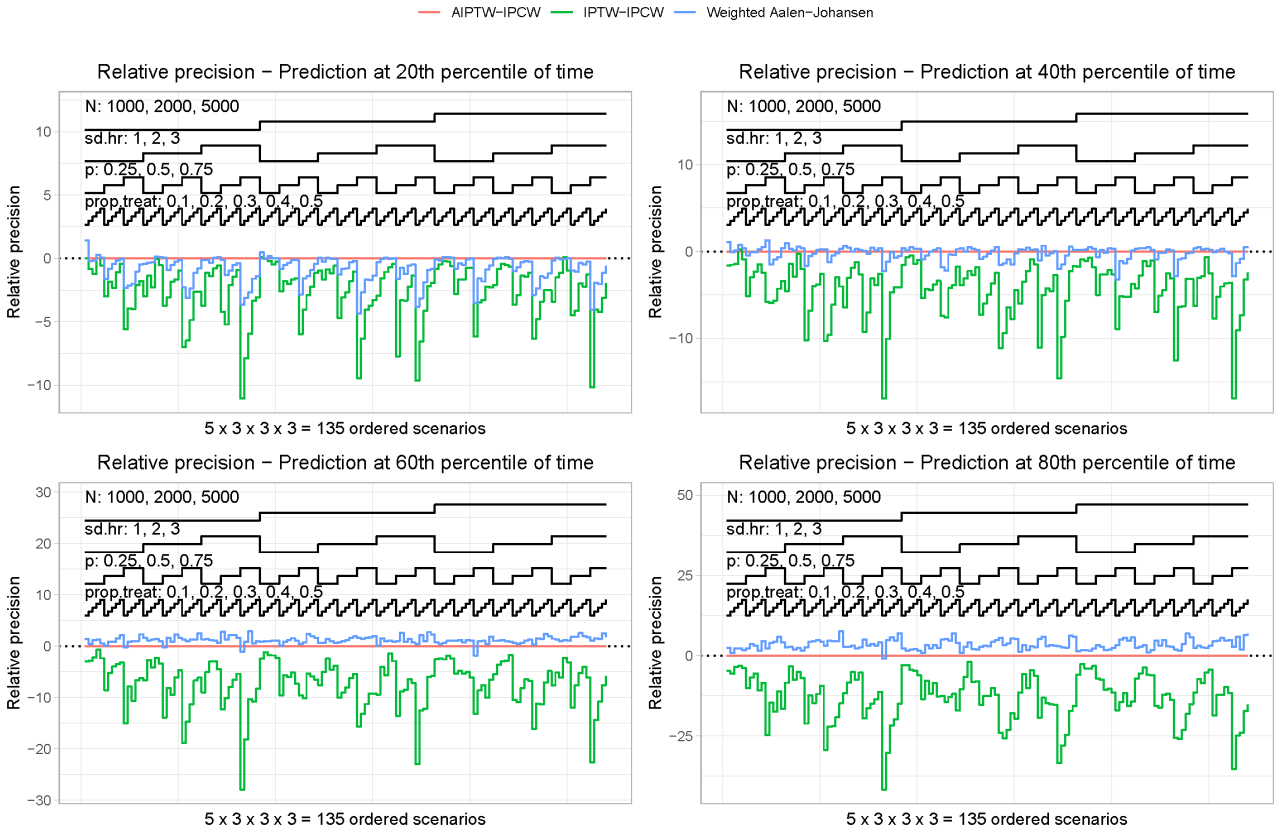

FIGURE 12 | Relative precision of estimated log−relative risk.

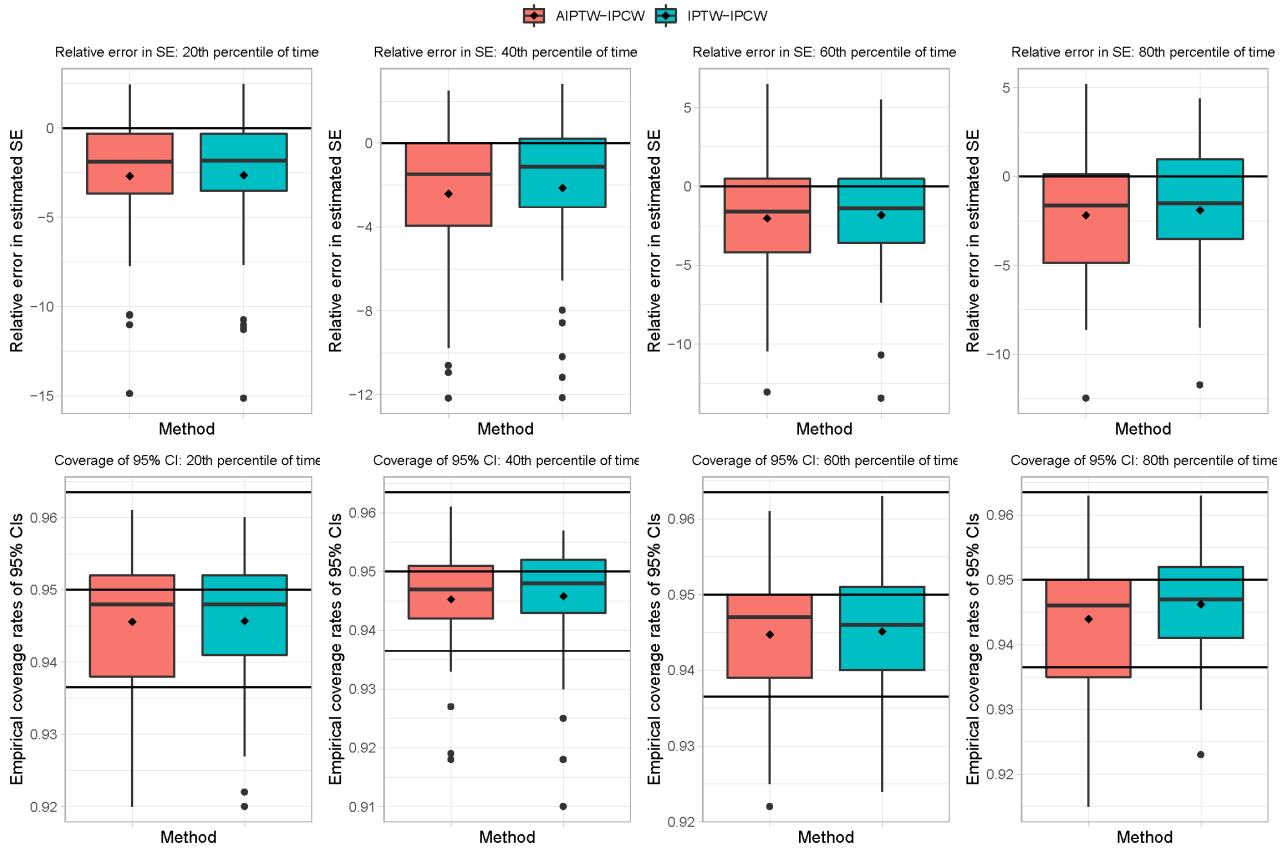

FIGURE 13 | Distribution of performance metrics for estimation of standard errors of relative risks.

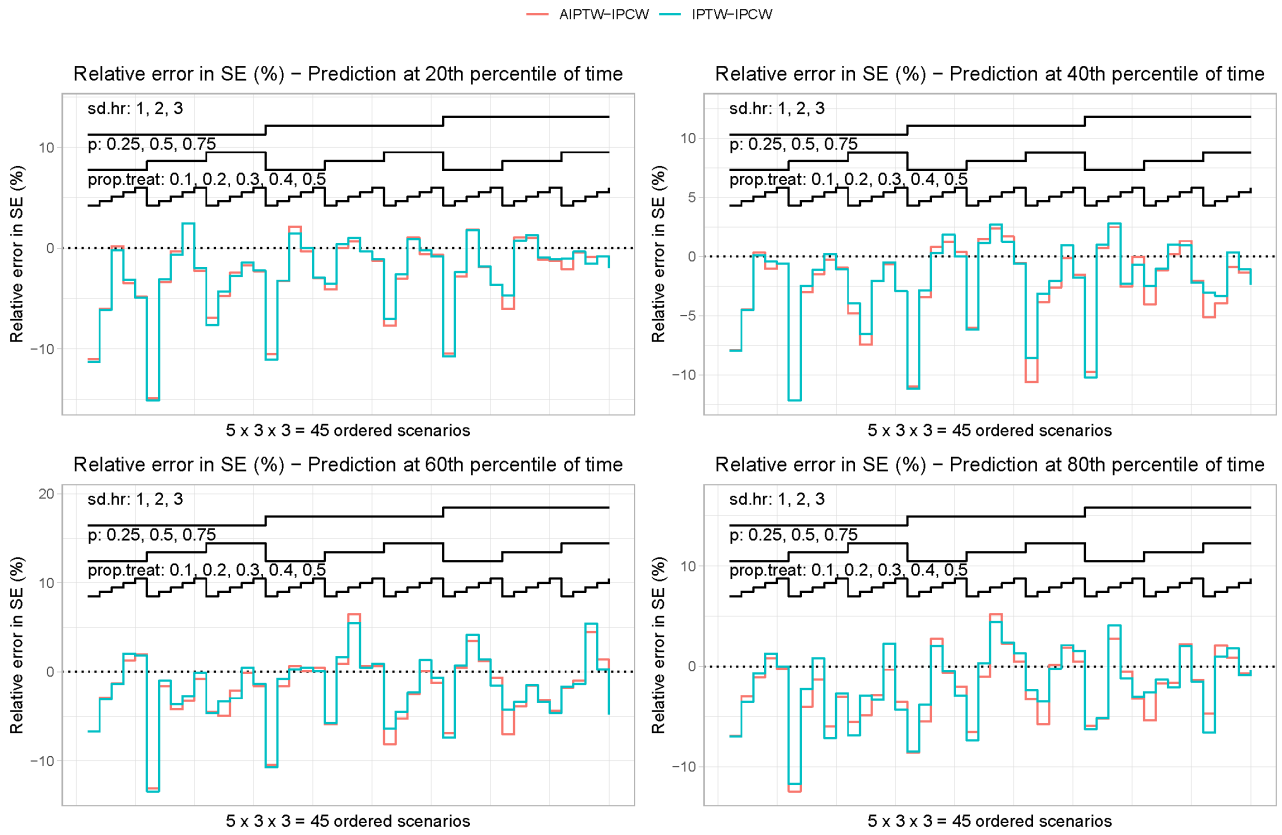

FIGURE 14 | Relative error in estimated standard error of estimated relative risk.

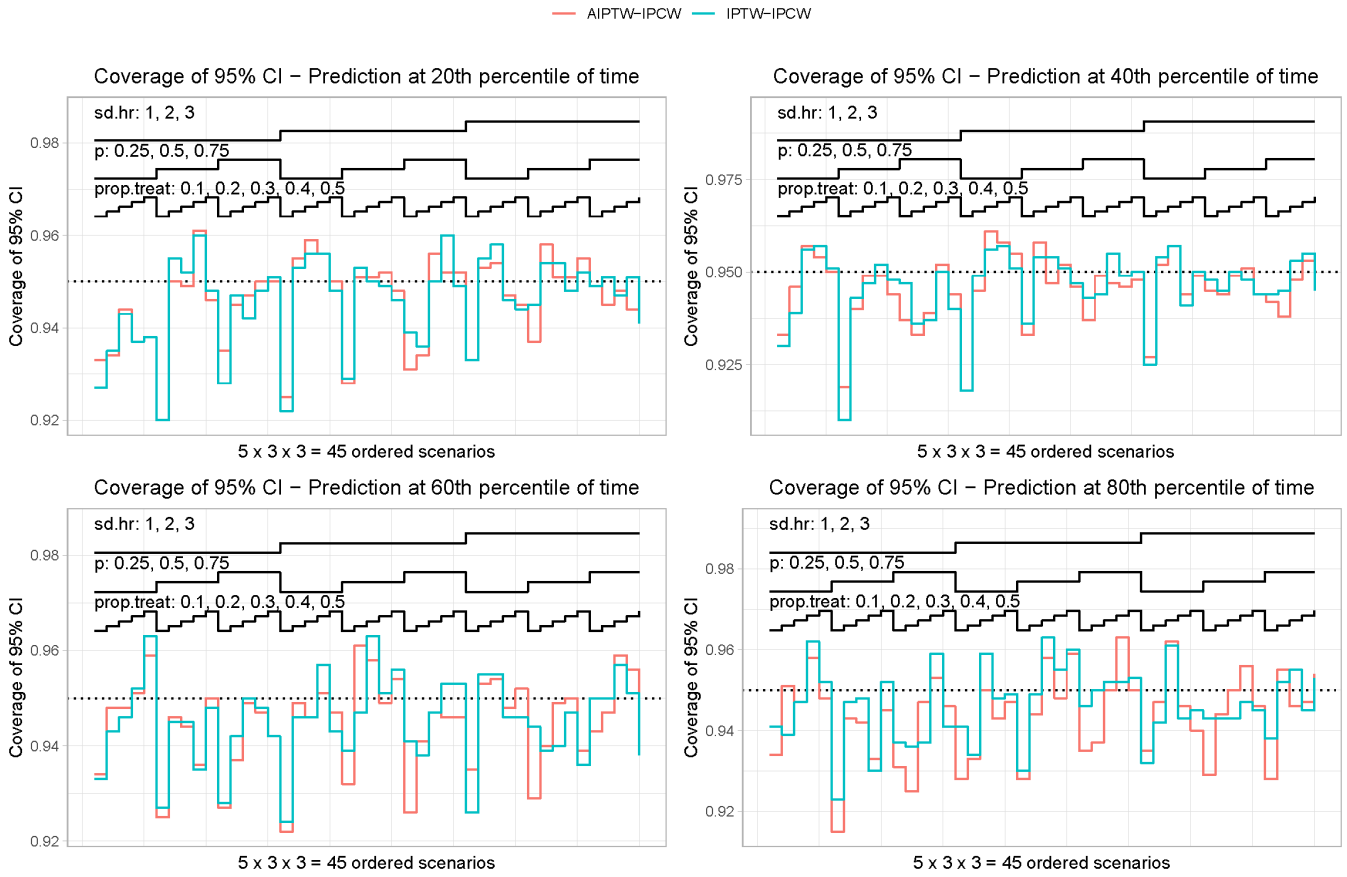

FIGURE 15 | Empirical coverage rates of 95% confidence intervals for estimated relative risks.

have added horizontal lines denoting these thresholds to the panels. In general, both methods tended to result in estimated confidence intervals whose empirical coverage rates did not differ from the advertised rate. When estimating relative risks at the 80th percentile of event times, IPTW-IPCW tended to have marginally better performance than AIPTW-IPCW. Nested loop plots comparing the performance of the two methods at each simulation scenario are reported in Figure 15.

7 | Discussion

The objective of the paper was three-fold: first, to describe methods to use IPTW to estimate the effects of treatments in settings with competing risks; second, to conduct empirical analyses illustrating the application of these methods; and third, to conduct Monte Carlo simulations to evaluate the relative performance of different methods for estimating time-specific risk differences and relative risks in settings with competing risks.

We suggest that applied researchers estimate both absolute and relative measures of effect. Absolute measures of treatment effect involve comparison in the differences in the probability of the occurrence of outcomes within specified durations of follow-up time. Absolute measures of effect can be complemented by the reporting of the NNT to avoid the occurrence of one outcome within a specified duration of time. Relative measures of effect can be either time-specific relative risks or marginal cause-specific hazard ratios. The recommendation to estimate the effect of treatment on both the cause-specific hazard and the CIF echoes the suggestion by Latouche et al. that "both hazards and cumulative incidence be analyzed side by side, and that this is generally the most rigorous scientific approach to analyzing competing risks data" [52].

In the current study, we have focused on estimating the effect of treatment on the primary outcome of interest. These analyses could be complemented by estimating the effect of treatment on the competing events. Estimating the effect of treatment on competing events may be helpful for identifying adverse consequences of the treatment. As noted above, there are alternative estimands, including the direct effect and the separable direct and indirect effects, that allow the analyst to address related questions.

We compared the performance of three estimators for time-specific risk differences and relative risks: the weighted Aalen–Johansen estimator, the IPTW-IPCW estimator, and the AIPTW-IPCW estimator. In the empirical analyses, the AIPTW-IPCW estimator failed to produce an estimate. In the simulations, we found that, while all three estimators tended to result in unbiased estimation of risk differences and relative risks, the IPTW-IPCW tended to result in estimates that were slightly less precise (i.e., had larger empirical standard errors) than the other two estimators. While the weighted Aalen–Johansen and AIPTW-IPCW estimators tended to have similar performance, there is an advantage to the weighted Aalen–Johansen estimator. The AIPTW-IPCW estimator, as

described by Ozenne and colleagues and as implemented in the ate() function in the riskRegression package for R, allows estimation of the ATE. Alternative sets of weights have been proposed that have different target estimands. These alternative sets of weights include average treatment effect in the treated (ATT) weights, matching weights, overlap weights, and entropy weights [53, 54]. An advantage of the weighted Aalen–Johansen estimator is that it can be easily modified to use any of these sets of weights. An advantage of the AIPTW-IPCW estimator is that it is doubly robust, meaning that the estimator will be consistent if at least one of the treatment-selection regression model and the outcomes regression model is specified correctly.

Throughout this study, we have assumed the use of parametric statistical models (e.g., logistic regression models) to estimate the treatment-selection model when computing the inverse probability of treatment weights. However, all the estimators that we considered can incorporate inverse probability of treatment weights derived using methods from the machine learning literature (e.g., random forests or stochastic gradient boosting machines). Similarly, we have focused on using nonparametric (e.g., Kaplan–Meier survival functions) or semi-parametric (e.g., Cox regression models) methods to model the censoring distribution when computing IPCWs. An alternative approach could be to use random survival forests to estimate these weights. Future research is warranted to examine the relative performance of different statistical and machine learning methods for estimating the different sets of weights.

In a previous study, we described how to use matching on the propensity score in the presence of competing risks [14]. Similarly to the current study, we recommended that analysts estimate both absolute and relative effects of treatment. In a propensity-score matched sample, CIFs can be estimated in treated and control subjects separately, allowing for the estimation of risk differences and relative risks at specified durations of time. This can be complemented by estimating a cause-specific hazard ratio by fitting a univariate cause-specific hazard model in the matched sample and using a robust variance estimator to account for the matched nature of the sample [4]. Matching on the propensity score has the ATT as the target estimand, while the weighting methods we have discussed have the ATE as the target estimand. Consequently, matching and weighting should not be seen as interchangeable.

The primary limitation of the current study is that our conclusions were based on Monte Carlo simulations. The design of these simulations was based on an analysis of empirical data, so that the simulations would reflect what was observed in a specific clinical setting. However, it is possible that different conclusions would be observed under a different data-generating process. The current study reflects much of the current research on the propensity score methods, in which simulations, rather than mathematical derivations, are employed [4, 6, 8, 9, 34, 55–68].

In conclusion, IPTW using the propensity score allows estimation of both the absolute and relative effects of treatments on outcomes in the presence of competing risks. Absolute treatment effects can be time-specific risk differences and the associated NNT. Relative effects of treatment can be cause-specific hazard ratios or time-specific relative risks.

Acknowledgments

ICES is an independent, nonprofit research institute funded by an annual grant from the Ontario Ministry of Health (MOH) and the Ministry of Long-Term Care (MLTC). As a prescribed entity under Ontario's privacy legislation, ICES is authorized to collect and use health care data for the purposes of health system analysis, evaluation, and decision support. Secure access to these data is governed by policies and procedures that are approved by the Information and Privacy Commissioner of Ontario. The use of the data in this project is authorized under Section 45 of Ontario's Personal Health Information Protection Act (PHIPA) and does not require review by a Research Ethics Board. This study was supported by ICES, which is funded by an annual grant from the Ontario MOH and the MLTC. This study also received funding from the Canadian Institutes of Health Research (CIHR) (PJT 183902). This document used data adapted from the Statistics Canada Postal CodeOM Conversion File, which is based on data licensed from Canada Post Corporation, and/or data adapted from the Ontario Ministry of Health Postal Code Conversion File, which contains data copied under license from Canada Post Corporation and Statistics Canada. Parts of this material are based on data and/or information compiled and provided by CIHI and the Ontario MOH. Parts of this report are based on Ontario Registrar General (ORG) information on deaths, the original source of which is Service Ontario. The views expressed therein are those of the author and do not necessarily reflect those of ORG or the Ministry of Public and Business Service Delivery. The analyses, conclusions, opinions, and statements expressed herein are solely those of the authors and do not reflect those of the funding or data sources; no endorsement is intended or should be inferred. The dataset from this study is held securely in coded form at ICES. While legal data-sharing agreements between ICES and data providers (e.g., healthcare organizations and government) prohibit ICES from making the dataset publicly available, access may be granted to those who meet pre-specified criteria for confidential access, available at www.ices.on.ca/DAS (email: das@ices.on.ca).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Research data are not shared.

References

1. P. R. Rosenbaum and D. B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70 (1983): 41–55.

2. P. C. Austin, "An Introduction to Propensity-Score Methods for Reducing the Effects of Confounding in Observational Studies," *Multivariate Behavioral Research* 46 (2011): 399–424.

3. P. C. Austin, "The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies," *Medical Decision Making* 29, no. 6 (2009): 661–677.

4. P. C. Austin, "The Performance of Different Propensity Score Methods for Estimating Marginal Hazard Ratios," *Statistics in Medicine* 32, no. 16 (2013): 2837–2849.

5. P. C. Austin, A. Manca, M. Zwarenstein, D. N. Juurlink, and M. B. Stanbrook, "A Substantial and Confusing Variation Exists in Handling of Baseline Covariates in Randomized Controlled Trials: A Review of Trials Published in Leading Medical Journals," *Journal of Clinical Epidemiology* 63, no. 2 (2010): 142–153.

6. P. C. Austin and T. Schuster, "The Performance of Different Propensity Score Methods for Estimating Absolute Effects of Treatments on Survival Outcomes: A Simulation Study," *Statistical Methods in Medical Research* 25, no. 5 (2016): 2214–2237.

7. 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.

8. E. Gayat, M. Resche-Rigon, J. Y. Mary, and R. Porcher, "Propensity Score Applied to Survival Data Analysis Through Proportional Hazards Models: A Monte Carlo Study," *Pharmaceutical Statistics* 11, no. 3 (2012): 222–229.

9. P. C. Austin, P. Grootendorst, S. L. Normand, and G. M. Anderson, "Conditioning on the Propensity Score Can Result in Biased Estimation of Common Measures of Treatment Effect: A Monte Carlo Study," *Statistics in Medicine* 26, no. 4 (2007): 754–768.

10. P. C. Austin, D. S. Lee, and J. P. Fine, "Introduction to the Analysis of Survival Data in the Presence of Competing Risks," *Circulation* 133 (2016): 601–609.

11. H. Putter, M. Fiocco, and R. B. Geskus, "Tutorial in Biostatistics: Competing Risks and Multi-State Models," *Statistics in Medicine* 26, no. 11 (2007): 2389–2430.

12. 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.

13. P. C. Austin, M. Ibrahim, and H. Putter, "Accounting for Competing Risks in Clinical Research," *Journal of the American Medical Association* 331, no. 24 (2024): 2125–2126.

14. P. C. Austin and J. P. Fine, "Propensity-Score Matching With Competing Risks in Survival Analysis," *Statistics in Medicine* 38, no. 5 (2019): 751–777.

15. S. R. Cole, B. Lau, J. J. Eron, et al., "Estimation of the Standardized Risk Difference and Ratio in a Competing Risks Framework: Application to Injection Drug Use and Progression to AIDS After Initiation of Antiretroviral Therapy," *American Journal of Epidemiology* 181, no. 4 (2015): 238–245.

16. S. R. Cole, M. G. Hudgens, M. A. Brookhart, and D. Westreich, "Risk," *American Journal of Epidemiology* 181, no. 4 (2015): 246–250.

17. B. M. H. Ozenne, T. H. Scheike, L. Staerk, and T. A. Gerds, "On the Estimation of Average Treatment Effects With Right-Censored Time to Event Outcome and Competing Risks," *Biometrical Journal* 62, no. 3 (2020): 751–763.

18. D. B. Rubin, "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66 (1974): 688–701.

19. J. G. Young, M. J. Stensrud, E. J. Tchetgen Tchetgen, and M. A. Hernan, "A Causal Framework for Classical Statistical Estimands in Failure-Time Settings With Competing Events," *Statistics in Medicine* 39, no. 8 (2020): 1199–1236.

20. M. J. Stensrud, M. A. Hernan, E. J. Tchetgen Tchetgen, J. M. Robins, V. Didelez, and J. G. Young, "A Generalized Theory of Separable Effects in Competing Event Settings," *Lifetime Data Analysis* 27 (2021): 588–631.

21. M. J. Stensrud, J. G. Young, V. Didelez, J. M. Robins, and M. A. Hernan, "Separable Effects for Causal Inference in the Presence of Competing Events," *Journal of the American Statistical Association* 117, no. 537 (2022): 175–183.

22. M. P. Fay and F. Li, "Causal Interpretation of the Hazard Ratio in Randomized Clinical Trials," *Clinical Trials* 21, no. 5 (2024): 623–635.

23. M. H. Gail, S. Wieand, and S. Piantadosi, "Biased Estimates of Treatment Effect in Randomized Experiments With Nonlinear Regressions and Omitted Covariates," *Biometrika* 7 (1984): 431–444.

24. P. C. Austin and A. Laupacis, "A Tutorial on Methods to Estimating Clinically and Policy-Meaningful Measures of Treatment Effects in Prospective Observational Studies: A Review," *International Journal of Biostatistics* 7, no. 1 (2011): 6–32.

25. H. F. Dorn, "Philosophy of Inference From Retrospective Studies," *American Journal of Public Health* 43 (1953): 692–699.

26. R. J. Cook and D. L. Sackett, "The Number Needed to Treat: A Clinically Useful Measure of Treatment Effect," *British Medical Journal* 310, no. 6977 (1995): 452–454.

27. A. Laupacis, D. L. Sackett, and R. S. Roberts, "An Assessment of Clinically Useful Measures of the Consequences of Treatment," *New England Journal of Medicine* 318 (1988): 1728–1733.

28. R. Jaeschke, G. Guyatt, H. Shannon, S. Walter, D. Cook, and N. Heddle, "Basic Statistics for Clinicians: 3. Assessing the Effects of Treatment: Measures of Association," *Canadian Medical Association Journal* 152, no. 3 (1995): 351–357.

29. E. Schechtman, "Odds Ratio, Relative Risk, Absolute Risk Reduction, and the Number Needed to Treat—Which of These Should We Use?," *Value in Health* 5 (2002): 431–436.

30. J. C. Sinclair and M. B. Bracken, "Clinically Useful Measures of Effect in Binary Analyses of Randomized Trials," *Journal of Clinical Epidemiology* 47, no. 8 (1994): 881–889.

31. 2018, http://www.bmj.com/about-bmj/resources-authors/article-types. accessed January 24, 2025.

32. D. G. Altman and P. K. Andersen, "Calculating the Number Needed to Treat for Trials Where the Outcome Is Time to an Event," *BMJ* 319, no. 7223 (1999): 1492–1495.

33. R. Sutradhar and P. C. Austin, "Relative Rates Not Relative Risks: Addressing a Wide-Spread Misinterpretation of Hazard Ratios," *Annals of Epidemiology* 28, no. 1 (2018): 54–57.

34. P. C. Austin, P. Grootendorst, and G. M. Anderson, "A Comparison of the Ability of Different Propensity Score Models to Balance Measured Variables Between Treated and Untreated Subjects: A Monte Carlo Study," *Statistics in Medicine* 26, no. 4 (2007): 734–753.

35. P. R. Rosenbaum, "Model-Based Direct Adjustment," *Journal of the American Statistical Association* 82 (1987): 387–394.

36. S. R. Cole and M. A. Hernan, "Adjusted Survival Curves With Inverse Probability Weights," *Computer Methods and Programs in Biomedicine* 75 (2004): 45–49.

37. M. A. Hernan, B. Brumback, and J. M. Robins, "Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men," *Epidemiology* 11, no. 5 (2000): 561–570.

38. P. C. Austin, "Variance Estimation When Using Inverse Probability of Treatment Weighting (IPTW) With Survival Analysis," *Statistics in Medicine* 35, no. 30 (2016): 5642–5655.

39. J. V. Tu, L. R. Donovan, D. S. Lee, et al., "Effectiveness of Public Report Cards for Improving the Quality of Cardiac Care: The EFFECT Study: A Randomized Trial," *Journal of the American Medical Association* 302, no. 21 (2009): 2330–2337.

40. J. V. Tu, A. Chu, L. R. Donovan, et al., "The Cardiovascular Health in Ambulatory Care Research Team (CANHEART): Using Big Data to Measure and Improve Cardiovascular Health and Healthcare Services," *Circulation. Cardiovascular Quality and Outcomes* 8, no. 2 (2015): 204–212.

41. K. A. Eagle, M. J. Lim, O. H. Dabbous, et al., "A Validated Prediction Model for all Forms of Acute Coronary Syndrome: Estimating the Risk of 6-Month Postdischarge Death in an International Registry," *Journal of the American Medical Association* 291, no. 22 (2004): 2727–2733.

42. F. E. Harrell, Jr., *Regression modeling strategies*, 2nd ed. (New York, NY: Springer-Verlag, 2015).

43. P. C. Austin and E. A. Stuart, "Moving Towards Best Practice When Using Inverse Probability of Treatment Weighting (IPTW) Using the Propensity Score to Estimate Causal Treatment Effects in Observational Studies," *Statistics in Medicine* 34 (2015): 3661–3679.

44. M. Mamdani, K. Sykora, P. Li, et al., "Reader's Guide to Critical Appraisal of Cohort Studies: 2. Assessing Potential for Confounding," *British Medical Journal* 330, no. 7497 (2005): 960–962.

45. P. C. Austin, A. Allignol, and J. P. Fine, "The Number of Primary Events per Variable Affects Estimation of the Subdistribution Hazard Competing Risks Model," *Journal of Clinical Epidemiology* 83 (2017): 75–84.

46. P. C. Austin, "The Iterative Bisection Procedure: A Useful Tool for Determining Parameter Values in Data-Generating Processes in Monte Carlo Simulations," *BMC Medical Research Methodology* 23, no. 1 (2023): 45.

47. J. Beyersmann, A. Allignol, and M. Schumacher, *Competing Risks and Multistate Models With R* (New York: Springer, 2012).

48. J. P. Fine and R. J. Gray, "A Proportional Hazards Model for the Subdistribution of a Competing Risk," *Journal of the American Statistical Association* 94 (1999): 496–509.

49. T. P. Morris, I. R. White, and M. J. Crowther, "Using Simulation Studies to Evaluate Statistical Methods," *Statistics in Medicine* 38, no. 11 (2019): 2074–2102.

50. *R: A Language and Environment for Statistical Computing [Computer Program]* (Vienna: R Foundation for Statistical Computing, 2005).

51. G. Rucker and G. Schwarzer, "Presenting Simulation Results in a Nested Loop Plot," *BMC Medical Research Methodology* 14 (2014): 129.

52. A. Latouche, A. Allignol, J. Beyersmann, M. Labopin, and J. P. Fine, "A Competing Risks Analysis Should Report Results on all Cause-Specific Hazards and Cumulative Incidence Functions," *Journal of Clinical Epidemiology* 66, no. 6 (2013): 648–653.

53. P. C. Austin, "Differences in Target Estimands Between Different Propensity Score-Based Weights," *Pharmacoepidemiology and Drug Safety* 32 (2023): 1103–1112.

54. Y. Zhou, R. A. Matsouaka, and L. Thomas, "Propensity Score Weighting Under Limited Overlap and Model Misspecification," *Statistical Methods in Medical Research* 29, no. 12 (2020): 3721–3756.

55. P. C. Austin, "Some Methods of Propensity-Score Matching Had Superior Performance to Others: Results of an Empirical Investigation and Monte Carlo Simulations," *Biometrical Journal* 51, no. 1 (2009): 171–184.

56. P. C. Austin, "Optimal Caliper Widths for Propensity-Score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies," *Pharmaceutical Statistics* 10 (2011): 150–161.

57. P. C. Austin, I. Type, and E. Rates, "Coverage of Confidence Intervals, and Variance Estimation in Propensity-Score Matched Analyses," *International Journal of Biostatistics* 5, no. 1 (2009): Article 13.

58. P. C. Austin, "The Performance of Different Propensity-Score Methods for Estimating Relative Risks," *Journal of Clinical Epidemiology* 61, no. 6 (2008): 537–545.

59. P. C. Austin, "The Performance of Different Propensity Score Methods for Estimating Marginal Odds Ratios," *Statistics in Medicine* 26, no. 16 (2007): 3078–3094.

60. P. C. Austin, "Statistical Criteria for Selecting the Optimal Number of Untreated Subjects Matched to Each Treated Subject When Using Many-To-One Matching on the Propensity Score," *American Journal of Epidemiology* 172, no. 9 (2010): 1092–1097.

61. P. C. Austin, "The Performance of Different Propensity-Score Methods for Estimating Differences in Proportions (Risk Differences or Absolute Risk Reductions) in Observational Studies," *Statistics in Medicine* 29, no. 20 (2010): 2137–2148.

62. P. C. Austin, "Comparing Paired vs Non-paired Statistical Methods of Analyses When Making Inferences About Absolute Risk Reductions in Propensity-Score Matched Samples," *Statistics in Medicine* 30, no. 11 (2011): 1292–1301.

63. P. C. Austin, "A Comparison of 12 Algorithms for Matching on the Propensity Score," *Statistics in Medicine* 33, no. 6 (2014): 1057–1069.

64. P. C. Austin and D. S. Small, "The Use of Bootstrapping When Using Propensity-Score Matching Without Replacement: A Simulation Study," *Statistics in Medicine* 33, no. 24 (2014): 4306–4319.

65. P. C. Austin, "Double Propensity-Score Adjustment: A Solution to Design Bias or Bias due to Incomplete Matching," *Statistical Methods in Medical Research* 26, no. 1 (2017): 201–222.

66. P. C. Austin and E. A. Stuart, "The Performance of Inverse Probability of Treatment Weighting and Full Matching on the Propensity Score in the Presence of Model Misspecification When Estimating the Effect of Treatment on Survival Outcomes," *Statistical Methods in Medical Research* 26, no. 4 (2017): 1654–1670.

67. P. C. Austin and E. A. Stuart, "Estimating the Effect of Treatment on Binary Outcomes Using Full Matching on the Propensity Score," *Statistical Methods in Medical Research* 26, no. 6 (2017): 2505–2525.

68. J. M. Franklin, W. Eddings, P. C. Austin, E. A. Stuart, and S. Schneeweiss, "Comparing the Performance of Propensity Score Methods in Healthcare Database Studies With Rare Outcomes," *Statistics in Medicine* 36, no. 12 (2017): 1946–1963.