Received: 17 March 2022 | Revised: 17 November 2022 | Accepted: 30 January 2023

DOI: 10.1002/sim.9681

RESEARCH ARTICLE

Statistics in Medicine WILEY

A comparison of different methods to adjust survival curves for confounders

Robin Denz | **Renate Klaaßen-Mielke** | **Nina Timmesfeld**

Department of Medical Informatics, Biometry and Epidemiology, Ruhr-University of Bochum, Bochum, North-Rhine Westphalia, Germany

Correspondence
Robin Denz, Department of Medical Informatics, Biometry and Epidemiology, Ruhr-University of Bochum, Universitätsstraße 105, D-44789, Bochum, North Rhine-Westphalia, Germany.
Email: robin.denz@rub.de

Treatment specific survival curves are an important tool to illustrate the treatment effect in studies with time-to-event outcomes. In non-randomized studies, unadjusted estimates can lead to biased depictions due to confounding. Multiple methods to adjust survival curves for confounders exist. However, it is currently unclear which method is the most appropriate in which situation. Our goal is to compare forms of inverse probability of treatment weighting, the G-Formula, propensity score matching, empirical likelihood estimation and augmented estimators as well as their pseudo-values based counterparts in different scenarios with a focus on their bias and goodness-of-fit. We provide a short review of all methods and illustrate their usage by contrasting the survival of smokers and non-smokers, using data from the German Epidemiological Trial on Ankle-Brachial-Index. Subsequently, we compare the methods using a Monte-Carlo simulation. We consider scenarios in which correctly or incorrectly specified models for describing the treatment assignment and the time-to-event outcome are used with varying sample sizes. The bias and goodness-of-fit is determined by taking the entire survival curve into account. When used properly, all methods showed no systematic bias in medium to large samples. Cox regression based methods, however, showed systematic bias in small samples. The goodness-of-fit varied greatly between different methods and scenarios. Methods utilizing an outcome model were more efficient than other techniques, while augmented estimators using an additional treatment assignment model were unbiased when either model was correct with a goodness-of-fit comparable to other methods. These "doubly-robust" methods have important advantages in every considered scenario.

KEYWORDS

adjusted survival curves, causal inference, confounding, simulation, time-to-event

1 | INTRODUCTION

In the analysis of clinical time-to-event data, treatment-specific survival curves are often used to graphically display the treatment effect in some population. The Kaplan-Meier estimator, stratified by treatment allocation, is usually used to calculate these curves. In sufficiently large randomized controlled trials with balanced groups this method yields

unbiased results.[1] In reality, it is often impossible to conduct such studies because of ethical or administrative reasons, which is why observational study designs are very common. As is well known, the lack of randomization can lead to the occurrence of confounding.[2,3] Simple Kaplan-Meier estimates do not take confounders into account and therefore produce a systematically biased picture of the true treatment effect in such cases.

The most popular way to adjust for confounders in medical time-to-event analysis is the use of the Cox proportional hazards model.[4] Communicating regression analysis results is, however, far more difficult than simply showing survival curves.[5-7] It has been shown multiple times that graphics simplify the communication of statistical results and lead to a better understanding by the reader.[8,9] Many researchers report both adjusted hazard-ratios and unadjusted Kaplan-Meier estimates.[10] Since the latter did not correct for the presence of confounders, these results often differ and hence confuse the reader.

Confounder-adjusted survival curves are a solution to this problem. Various methods for calculating these have been developed.[11-17] Their properties have only been studied to a limited extent. Theoretical results can only be generalized to a certain degree because some assumptions are quite complex. Previous simulation studies deal exclusively with the properties of the survival curves at specific points in time.[14,18-20] Our main goal is to fill this gap by investigating the properties of the available methods with respect to the entire survival curve using a Monte-Carlo simulation. We aim to determine which of these methods produce unbiased estimates and which methods show the least deviation from the true survival curve on average (goodness-of-fit).

This article is structured as follows. First, we give a formal description of confounder-adjusted survival curves and the background. Afterwards, a brief description of all included methods is given. Using real data from a large prospective cohort study, we illustrate the usage of these methods by comparing the survival of non-smokers and current or past smokers. Next, the design of the simulation study is described and the results are presented. Finally, we discuss the results and their implications for the practical applications of the adjustment methods.

2 | BACKGROUND AND NOTATION

Let $Z \in \{0, 1, \dots, k\}$ denote the treatment group, where each value of Z indicates one of k possible treatments. For the sake of clarity, we use the term "treatment groups" throughout the article, but it is important to note that any other grouping variable may be used as well. Let T be the time to the occurrence of the event of interest. In reality, it is sometimes only known whether a person has suffered an event by time C or not, which is known as right-censored data. In this case, only $T_{obs} = \min(T, C)$ would be observed with a corresponding event indicator $D = I(T < C)$. Although a crucial point for the estimation methods, it is unimportant for the definition of the target estimand.

Under the Neyman-Rubin causal framework, every person has k potential survival times $T^z \in \{T^0, T^1, \dots, T^k\}$, one for each of the k possible treatment strategies.[2,21] The goal is to estimate the counterfactual survival probability in the target population over time, where every person has received the same treatment. This population consists of N individuals, indexed by i, $i = 1, 2, \dots, N$, each with their own vector of baseline covariates x_i. The counterfactual survival probability of individual i at time t is defined by:

$$S(t|Z = z, X = x_i) = P(T^z > t|x_i), \tag{1}$$

where T^z denotes the failure time which would have been observed, if treatment $Z = z$ was actually administered.[19] Therefore, the target function is defined as:

$$S_z(t) = E(I(T^z > t)). \tag{2}$$

In the literature, this quantity is often called the *causal survival curve*, the *counterfactual survival curve* or the *confounder-adjusted survival curve*. We use all of these terms interchangeably. It represents the survival probability that would be observed in the target-population, if every person in the population had received treatment Z. The difference or ratio between two treatment-specific counterfactual survival curves is sometimes used to define the *average treatment effect*.[14] Since $S_z(t)$ refers to the entire target-population, irrespective of the observed treatment status, it may also be considered an average treatment effect by itself. Randomized controlled trials are the gold standard for their estimation. While it is fundamentally impossible to observe more than one potential survival time at once, randomization ensures

that the distribution of all other variables does not differ between the treatment groups on average. Therefore, differences between the groups can only be attributed to the treatment itself.[2,3]

In order to estimate such an effect without randomization, three assumptions have to be met: the *stable unit treatment value assumption* (the potential survival time of one person is independent of the treatment assignment of other people in the study), the *no unmeasured confounding assumption* (all relevant confounders have been measured) and the *positivity assumption* (every person has a probability greater than 0 and smaller than 1 for receiving treatment z). Those are described in detail elsewhere.[3,22-24] Although the methods discussed in this article are vastly different, they all share these fundamental assumptions. If any one of them is violated, $S_z(t)$ is not identifiable.

3 | OVERVIEW OF METHODS

In this article, we focus strictly on methods that can be used to adjust survival curves for measured baseline confounders, when random right-censoring is present. Methods which are concerned with covariate adjustment in order to increase statistical power only,[25] corrections for covariate-dependent censoring,[26] time-varying confounding,[27] and unmeasured confounders[28] are disregarded. The *Average Covariate* method, which entails fitting a Cox model to the data and plugging in the mean of all covariates in order to predict the survival probability for each treatment at a range of time points,[29] is also discarded. We choose to do this, because it has been shown repeatedly that this method produces biased estimates.[30,31] Multiple methods based on stratification[30,32-34] are also excluded, because they are only defined for categorical confounders. Additionally, we choose to exclude *Targeted Maximum Likelihood Estimation* based methods,[19,35,36] because they are currently only defined for discrete-time survival data. Artificial discretization of continuous-time survival data is problematic theoretically[37] and has been shown to not work very well in practice.[38,39] Current implementations of these methods are also very computationally complex, making it infeasible to include them in our simulation study.

According to these constraints, the following adjustment methods are included in our simulation study: the *G-Formula*, *Inverse Probability of Treatment Weighting*, *Propensity Score Matching*, *Empirical Likelihood Estimation*, *Augmented Inverse Probability of Treatment Weighting* and their *Pseudo Values* based counterparts. A standard Kaplan-Meier estimator is also included to illustrate the impact of missing confounder-adjustment. An exhaustive description of each method is beyond the scope of this article, but a short review of each included method is given below. More details can be found in the cited literature. Table 1 includes a brief summary of the most important aspects of each method and the Online Appendix includes a step-by-step guide to each method presented here.

TABLE 1 Summary of the qualitative properties of each method.

Method	Requires an outcome model	Requires a treatment model	Doubly robust	More than two treatments allowed	Monotone estimates	Bounded estimates	SE formula
G-Formula	✓	×	×	(✓)[a]	(✓)[a]	(✓)[a]	✓
G-Formula PV	✓	×	×	✓	×	✓	×
IPTW KM	×	✓	×	✓	✓	✓	✓
IPTW HZ	×	✓	×	✓	✓	✓	×
IPTW PV	×	✓	×	✓	×	×	✓
Matching	×	✓	×	✓	✓	✓	×
EL	×	×	✓	(×)[b]	✓	✓	×
AIPTW	✓	✓	✓	✓	×	×	✓
AIPTW PV	✓	✓	✓	✓	×	×	✓
G-Formula IPTW	✓	✓	?[c]	(✓)[a]	(✓)[a]	(✓)[a]	×

[a]This is true when using a Cox model, but might vary otherwise.
[b]It has not been proposed in the literature, but can be extended to this case fairly easily.
[c]This property is currently unclear.

All methods mentioned above can be roughly divided into three categories: methods utilizing the *outcome mechanism*, methods that use the *treatment assignment mechanism* and methods relying on *both* types of mechanisms. In survival analysis, using the outcome mechanism refers to modeling the process which determines the time until the event of interest occurs. Formally, the goal is to use a statistical model to obtain a plug-in estimator of $S(t|Z, X)$ (see Equation 1). This model has to take right-censoring into account. The treatment assignment mechanism, on the other hand, describes the process by which an individual i is assigned to one of the k possible treatments. The goal is to estimate the probability of receiving treatment z for each individual, denoted by $P(Z = z|X)$, which is formally known as the *propensity score*.[22,40] A statistical model is usually necessary to estimate this probability. How exactly the outcome and treatment models are used to calculate the causal survival curves is described below.

3.1 | G-Formula

One well known method to adjust for confounders is the G-Formula, also known as G-computation, direct standardization, or corrected-group-prognosis method in epidemiology.[11,29,41] Here, the confounders are adjusted for by correctly modeling the outcome mechanism. The Cox proportional hazards model[4] in conjunction with an estimate of the baseline-hazard function[42] is usually used. However, any model allowing predictions for the conditional survival probability, given covariates and a point in time, may be used as well.[41] After the model has been estimated, it is used to make predictions for $S(t|Z = z, X = x_i)$ under each possible treatment for each individual. If the estimated conditional survival probabilities are unbiased, the resulting arithmetic mean of these predictions over all individuals is an unbiased estimate for the counterfactual survival probability at time t.

For survival data, this method was first proposed independently by Makuch[11] and Chang et al,[29] using a simple Cox model. Other authors have used different models as well,[16,43] but we will only consider the Cox model in this article. This method has been shown to outperform other methods in terms of efficiency, especially when strong predictors of the outcome are included in the model.[20,44]

3.2 | Inverse probability of treatment weighting (IPTW)

Inverse probability of treatment weighting (IPTW) is one of multiple methods utilizing the treatment assignment mechanism for confounder-adjustment.[12,45] First, the propensity scores are estimated for each individual and each treatment. Afterwards, the inverse of this probability is calculated. By using these values as weights in the analysis, the confounding is removed and unbiased estimates of the actual causal effect can be obtained. In practice, the propensity score is usually estimated using a logistic regression model, but other more sophisticated methods exist as well.[46,47]

For survival analysis, two very similar estimators of the causal survival curve based on this method have been proposed.[12,45] The estimator of Cole and Hernán[12] (IPTW HZ) is equivalent to fitting a weighted stratified Cox model, using the treatment indicator as stratification variable. The inverse probability weights are used as case weights. Xie and Liu[45] on the other hand proposed a directly weighted Kaplan-Meier estimator (IPTW KM).

Previous research on IPTW in the estimation of sample means has shown repeatedly that IPTW is less efficient than the G-Formula when both methods are used appropriately.[48] Efficiency is judged by the size of the SEs in this context. This difference is particularly strong when the estimated weights are highly variable, close to 0 or extremely high, which often happens in small sample sizes.[49] Similar results have recently been reported for point estimates of the survival curve.[20]

3.3 | Propensity score matching

Another well known method is propensity score matching. Instead of using the propensity scores to construct weights for each observation, a new sample is constructed by matching patients with similar propensity scores. The new matched sample can subsequently be analysed using standard methods. Slightly different methods for creating adjusted survival curves from matched data have been proposed.[13,50-52] We roughly follow the method described in Austin[13] and use the following three steps: (1) estimation of the propensity score, (2) matching patients with similar scores, and (3) using a

simple stratified Kaplan-Meier estimator on the matched sample. We used the algorithm of Sekhon[53] with an Euclidean distance and allowing both replacement and ties. If the propensity scores are correctly estimated and the algorithm used for matching the patients is appropriate, the resulting estimates of the survival curves are unbiased.[13] Propensity score matching has, however, been shown to be less efficient than IPTW and the G-Formula in other aspects of the analysis of time-to-event data.[54,55]

3.4 | Augmented inverse probability of treatment weighting (AIPTW)

A different approach, called locally efficient augmented inverse probability weighting (AIPTW), was first proposed by Robins and Rotnitzky[56] and Hubbard et al[57] and further developed by different authors in the present context.[15,44,58] Instead of using a single treatment assignment or outcome model, this method requires *both* kinds of models. In simplified terms, the AIPTW methodology works by using the G-Formula estimate to augment the IPTW estimate, in order to make it more efficient. Essentially, it is just the IPTW estimator with the conditional survival predictions under each treatment added to it, after weighting them using the propensity score. The specific form of the estimating equation ensures that it is asymptotically unbiased if *either* of the two models is correctly specified, making it *doubly-robust*. This property is the main advantage of this method.

It has been shown previously that this method is asymptotically at least as efficient as IPTW, when both models are correctly specified.[44,58] Research on using AIPTW for the estimation of biased sample means also indicates, however, that the method looses efficiency if both models are slightly incorrectly specified.[59] Because point estimates are made without any global constraints, there is no guarantee that the estimates will be monotonically decreasing, or that the estimates will be between 0 and 1. How often these problems occur in practice is currently not clear.

3.5 | G-Formula + IPTW

Another possible way to combine the G-Formula estimator with the IPTW approach has recently been proposed by Chatton et al,[16] based on earlier work by Vansteelandt and Keiding.[60] This method works as follows. First, the inverse probability of treatment weights are estimated. Those weights are then used in the estimation of an outcome model. The outcome model should also include the relevant confounders as covariates. This model can then be used to calculate standard G-Formula estimates as described in section 3.1.[16] As in the usual G-Formula approach, any outcome model could theoretically be used, but we only consider a Cox model in this article.

Although Chatton et al[16] performed a well designed Monte-Carlo simulation, where this method showed the doubly-robust property when estimating the hazard-ratio, they offer no mathematical proof. In addition, they did not study the estimation of the adjusted survival curve directly. Instead they focused on two other useful quantities, the hazard-ratio and the restricted mean survival time. It is unclear whether their results extend to the given context. We therefore use the name *G-Formula IPTW* instead of *Doubly-Robust Standardization*, which is the name given by the original authors.

3.6 | Empirical likelihood estimation (EL)

Recently, Wang et al[14] have proposed an estimator of the survival function, which is based on the empirical likelihood (EL) estimation methodology. Stated simply, EL is a likelihood based method, which does not require the assumption that the data was generated by any known family of distributions. It is a model free approach that works by forcing the moments of the covariates X to be equal between treatment groups, through the maximization of a constrained likelihood function. The resulting equality of the distributions removes the bias created by the confounders. No specification of either outcome model or treatment allocation model is necessary. Due to the constraints in the optimization process, this method is guaranteed to produce non-increasing estimates lying in the probability bounds.

Theoretical results and simulation studies of Wang et al[14] and Lee et al[61] indicate that this method also shares the doubly-robust property. It has been demonstrated previously, that EL can outperform IPTW in terms of variance in some

scenarios.[14,62] The method described in Wang et al[14] itself is less efficient than the standard AIPTW estimator, but can be modified to increase its efficiency using the method described by Lee et al.[61] In this article, we focus strictly on the version described in the original article by Wang et al[14] because of code availability.

3.7 | Pseudo values (PV)

An alternative approach is the use of pseudo values (PV).[17,63] The general idea is that one could use standard methods, such as generalized linear models, to model the outcome mechanism, if T was known for each individual i. As described in Section 2 this is usually not the case in practice. Generally, only T_{obs} is observed because of right-censoring. Let $\hat{S}(t)$ be the standard Kaplan-Meier estimator. The PV for individual i at time t is then defined as:

$$\hat{\theta}_i(t) = n\hat{S}(t) - (n-1)\hat{S}^{-i}(t), \tag{3}$$

where $\hat{S}^{-i}(t)$ is the Kaplan-Meier estimator applied to the data of size $n-1$, where the ith observation has been removed. The PV for individual i can be interpreted as the contribution of the individual i to the target estimate from a complete sample of size n without censoring.[63,64]

After the PVs are estimated for each person at a fixed set of points in time, they can be used to construct G-Formula, IPTW or AIPTW estimates of the survival curve.[17,18,64] For the G-Formula estimate, a generalized estimating equation (GEE) model can be fit, using the PVs as response variable and all baseline covariates as independent variables for each time point. This model can then be used to obtain the required conditional survival probability predictions.[65,66] For the IPTW estimate, a simple propensity score weighted average of the PVs can be used.[17] The AIPTW estimate based on PV can be obtained in a similar fashion.[18,64] Little is known about the efficiency of these estimators compared to the non-PV based counterparts. Regardless which method based on PV is used, there is no guarantee that the survival probability will be monotonically decreasing over time. Isotonic regression can be used to correct these errors after estimating the curves.[67] The impact of using this correction method will also be assessed in the simulation study.

4 | ILLUSTRATIVE EXAMPLE

To illustrate the use of the methods described above we consider the estimation of counterfactual survival curves of smokers vs non-smokers. It is well known that smoking tobacco increases the probability of developing certain types of cancer and a multitude of other illnesses, thereby reducing the average life expectancy of smokers.[68,69] The total causal effect of smoking on the survival time is however difficult to assess, because it is impossible to conduct randomized trials due to ethical and administrative reasons. This has lead to a lot of discussion about confounding factors, especially in the late 1950s and early 1960s when the biological mechanisms that link smoking to a higher risk of cancer were not well known.[70] By now it is known that age,[71] gender[72] and the level of education[73] are some of the confounders that need to be accounted for when estimating the total causal effect of smoking on the survival time. These factors tend to increase the propensity to smoke and decrease the survival time. Crude Kaplan-Meier curves stratified by smoking status therefore usually show an *overestimate* of the true causal effect, meaning that the curves are too far apart.

We illustrate this using data from the German Epidemiological Trial on Ankle-Brachial-Index (getABI), which is a prospective observational cohort study including a total of 6880 primary care patients aged 65 or older. At the beginning of the trial in 2001, an examination was performed on every included patient. Data on both smoking status and the relevant confounders listed above was collected during this examination. In total, 3687 patients reported to have never smoked and 3134 patients had smoked in the past or are current smokers. Follow-up information on these patients was collected for up to 7 years after the baseline examination. Detailed information about the design of the trial and the obtained results can be found elsewhere.[74,75]

Figure 1 shows the crude and adjusted survival curves for people who have never smoked vs people who are current smokers or have smoked in the past. We used every method mentioned in Section 3 to adjust the curves for age, gender, and educational status. We used a Cox regression to model the survival time and a logistic regression to model the treatment assignment. Regardless of the method, it can clearly be seen that the adjusted survival curves are closer to each other

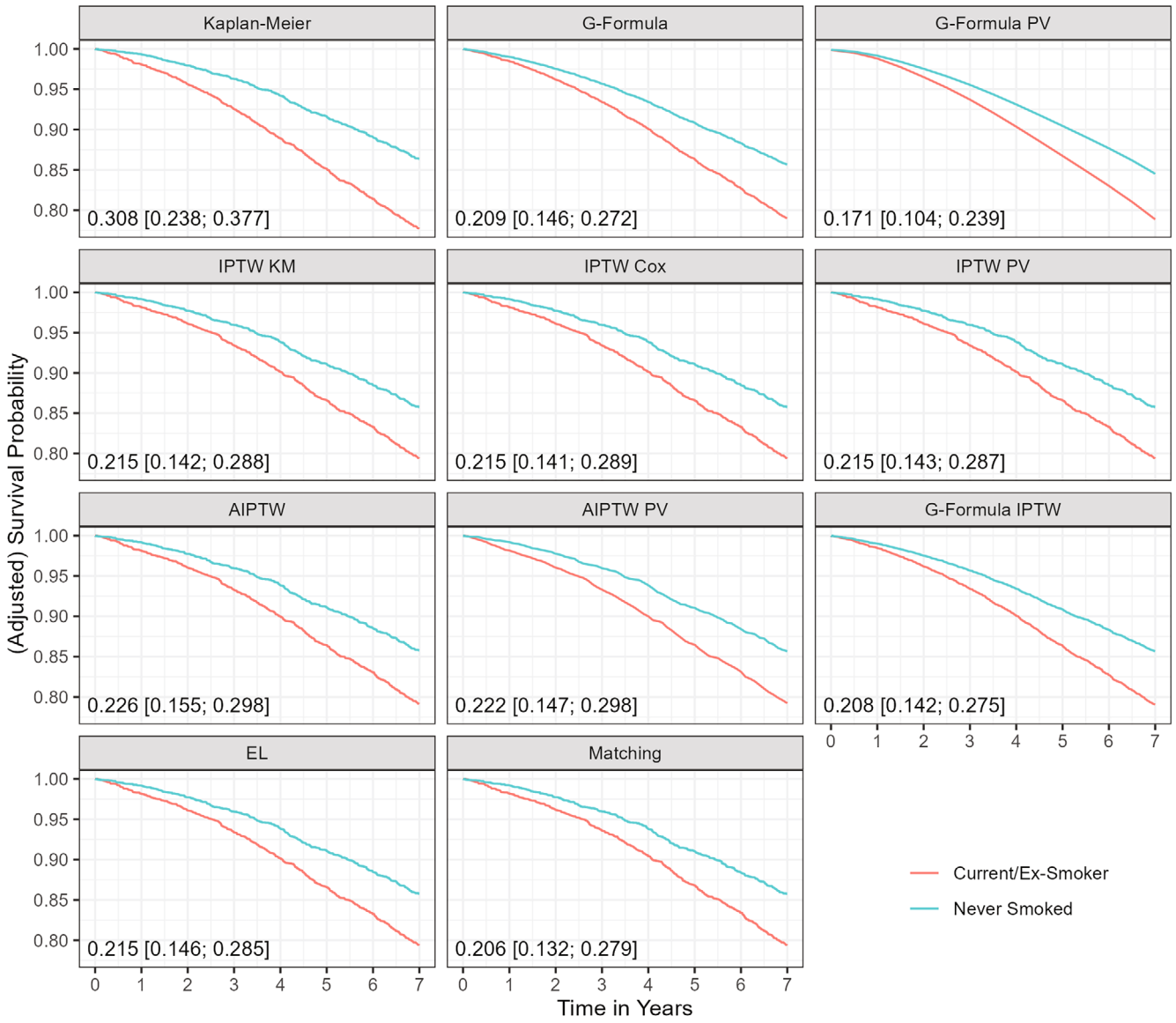

FIGURE 1 Crude and adjusted survival curves for people who smoked in the past or smoke currently vs people that never smoked. The numbers in the left corner of each facet are the area between the curves in the interval [0, 7] years and its associated 95% bootstrap confidence interval, calculated using 300 bootstrap replications and numerical integration.

than the crude Kaplan-Meier curves. Nonetheless, the survival curve of never-smokers is still above smokers in the entire interval, suggesting that the effect of smoking on the survival time is not entirely due to the considered confounders. The differences between the adjustment methods are rather small in this case. Additional information about the estimation of these curves, including checks of the proportional hazards assumption and positivity violations are given in the Online Appendix.

5 | SIMULATION STUDY

5.1 | Data generation and procedure

The simulation is designed as a mixture of the approaches of Austin et al[52] and Chatton et al.[20] A causal diagram showing all causal pathways and associated coefficients is displayed in Figure 2. The following five steps are executed:

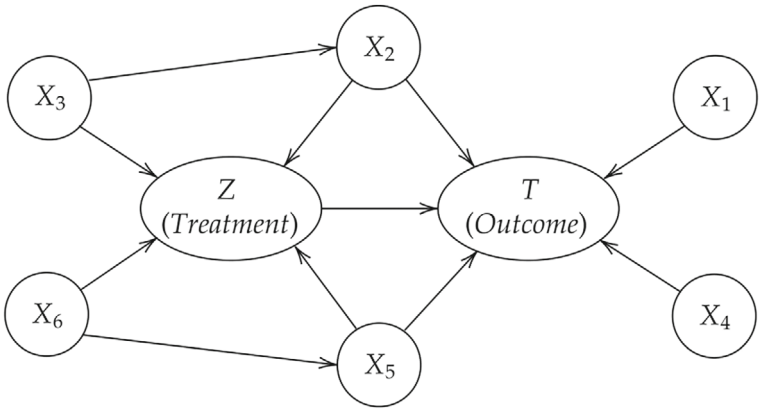

FIGURE 2 Causal diagram of the data generation mechanism used in the simulation. *Source*: Adapted from Chatton et al.[20]

Step 1 First, a *super-population* of one million people is generated. The distributions and causal coefficients used are listed below.

A. Four independent and identically distributed variables (X_1, X_3, X_4, X_6) are generated for each person.
B. Two further variables (X_2, X_5) are generated, where X_2 is partially caused by X_3 and X_5 is partially caused by X_6.
C. Based on those six variables and the causal coefficients listed below two survival times are simulated for each individual (using Z, X_1, X_2, X_4 with linear effects and X_5 with a quadratic effect). One for $Z = 0$ (control group) and one for $Z = 1$ (treatment group), using the same stream of random numbers. The result is a population of individuals in which both potential survival times are known. To generate the survival times we use the method of Bender et al[76] with a Weibull distribution ($\lambda = 2, \gamma = 1.8$).
D. Following Chatton et al,[20] the probability of receiving the treatment $P(Z = 1|X)$ is determined by a logistic regression model, using the covariates X_2, X_3, X_5, and X_6 and the causal coefficients listed below, where X_2 has a quadratic effect on the probability. The intercept is set to -1.2, which results in approximately 50% of all individuals receiving the treatment.
E. The true counterfactual survival curves $S_z(t)$ are calculated using simple proportions:

$$S_z(t) = \frac{1}{1000000} \sum_{i}^{1000000} I(t_{iz} > t), \tag{4}$$

where t_{iz} is the potential survival time of individual i given treatment z. Those are displayed in the Online Appendix.

Step 2 A simple random sample without replacement is drawn from the super-population.

Step 3 For each person in the sample, the treatment status z_i is generated using a Bernoulli trial using the previously estimated probability of receiving treatment. Only the survival time corresponding to the drawn treatment status is kept in the sample.

Step 4 Random right-censoring is introduced using another Weibull distribution ($\lambda = 1, \gamma = 2$), resulting in 22.90% of right-censored individuals on average.

Step 5 All considered methods are used to estimate the treatment-specific causal survival probability for each treatment level at a fine grid of points in time (37 equally spaced points from 0.05 to 1.5).

Steps two to five are repeated numerous times in different simulation scenarios, which are described in more detail below. Since the probability of treatment allocation is dependent on the variables X_2 and X_5, which also have a direct causal effect on the outcome, the causal effect of the treatment on the survival time is confounded. Failing to adjust for this confounding would lead to biased estimates of the treatment specific survival curves. The causal structure of the samples created by carrying out the procedure described above is displayed in Figure 2.

We used the following distributions and causal coefficients to generate the samples:

$$
\begin{aligned}
X_1 &\sim Bernoulli(0.5)\\
X_2 &\sim Bernoulli(0.3 + X_3 \cdot 0.1)\\
X_3 &\sim Bernoulli(0.5)\\
X_4 &\sim N(0,1)\\
X_5 &\sim 0.3 + X_6 \cdot 0.1 + N(0,1)\\
X_6 &\sim N(0,1)\\
Z &\sim Bernoulli\left(\frac{1}{1+\exp(-(-1.2+\log(3)\cdot X_2^2+\log(1.5)\cdot X_3+\log(1.5)\cdot X_5+\log(2)\cdot X_6))}\right)\\
T &\sim \left(-\frac{\log(U(0,1))}{\exp(\log(1.8)\cdot X_1+\log(1.8)\cdot X_2+\log(1.8)\cdot X_4+\log(2.3)\cdot X_5^2-1\cdot Z)}\right)^{0.5},
\end{aligned}
$$

where *Bernoulli*() corresponds to simple Bernoulli trials, $N(0,1)$ is the standard normal distribution and $U(0,1)$ is a uniform distribution of numbers between 0 and 1. All calculations were performed using the **R** programming language (version 4.2.1) with the `adjustedCurves` package available on CRAN,[77] which was developed by the authors to facilitate the use of adjusted survival curves in practice. The code used is available in the Online Appendix.

5.2 | Scenarios

We consider five main scenarios, in which different correctly and incorrectly specified models for the treatment assignment and outcome mechanism are used. In all of these scenarios the variables X_1, X_2, X_4, X_5, and Z are used as independent variables to model the outcome mechanism and the variables X_2 and X_5 are used to model the treatment-assignment mechanism. We choose to include all outcome predictors in the outcome model to prevent the problem of unobserved heterogeneity. If, for example, X_1 was omitted from the model, this could lead to a violation of the proportional hazards assumption when using the Cox model and therefore lead to biased estimates, even if the confounders X_2 and X_5 were correctly adjusted for.[78] The only difference between the scenarios is whether the quadratic effects of X_5 on the outcome and of X_2 on the treatment-assignment was specified correctly:

1. **Correct outcome mechanism and correct treatment assignment** (CO & CT): X_2 is modeled as a quadratic effect on the treatment-assignment and as a linear effect on the outcome. X_5 is modeled as a linear effect on the treatment-assignment and as a quadratic effect on the outcome.
2. **Correct outcome mechanism and incorrect treatment assignment** (CO & ICT): X_2 is modeled as a linear effect on both the treatment-assignment and the outcome. X_5 is modeled as a quadratic effect on both the treatment-assignment and the outcome.
3. **Incorrect outcome mechanism and correct treatment assignment** (ICO & CT): X_2 is modeled as a quadratic effect on both the treatment-assignment and the outcome. X_5 is modeled as a linear effect on both the treatment-assignment and the outcome.
4. **Incorrect outcome mechanism and incorrect treatment assignment** (ICO & ICT): X_2 is modeled as a linear effect on the treatment-assignment and as a quadratic effect on the outcome. X_5 is modeled as a quadratic effect on the treatment-assignment and as a linear effect on the outcome.
5. **Partially correct outcome mechanism and partially correct treatment assignment** (PCO & PCT): Both X_2 and X_5 are modeled as linear effects in both models.

Simple logistic regression models are used for methods relying on a treatment assignment model (IPTW KM, IPTW HZ, IPTW PV, PS Matching, AIPTW, AIPTW PV, G-Formula IPTW). Cox models are used for methods relying on an outcome model which takes right-censoring into account (G-Formula, AIPTW, G-Formula IPTW) and a GEE model is used in the corresponding PV based methods (G-Formula PV, AIPTW PV). The EL method does not utilize any

models and therefore only received the corresponding raw design matrix as input. For example, in the CO & CT scenario the EL method received a matrix containing the variables $X_1, X_2, X_2^2, X_4, X_5, X_5^2$ whereas it only received the variables $X_1, X_2, X_4, X_5, X_5^2$ in scenario CO & ICT. For all five of the scenarios mentioned above the sample size is varied systematically. In doing so, a broad range of realistic scenarios is covered, allowing detailed judgments about the performance of each method. The Online Appendix additionally includes further simulation scenarios, in which the covariate sets used were varied.

5.3 | Performance criteria

All estimates are compared to the true survival curves. In contrast to previous studies, our main interest is to estimate the performance of the estimators concerning the entire survival curves.[14,18-20] Usually, the bias is defined as $E(S_z(t) - \hat{S}_z(t))$ for a particular point in time t. We remove this time dependency by defining the generalized bias in group z as:

$$G_{\text{Bias}}(z) = \int_0^\infty E\left(S_z(t) - \hat{S}_z(t)\right) dt = E\left(\int_0^\infty S_z(t) - \hat{S}_z(t) dt\right), \tag{5}$$

where $\hat{S}_z(t)$ is the estimated survival function for treatment $Z = z$ and $S_z(t)$ is the true survival function for $Z = z$. To estimate the integral for a particular simulation run we use:

$$\hat{\Delta}_{\text{Bias}}(z) = \int_0^\tau \left(S_z(t) - \hat{S}_z(t)\right) dt, \tag{6}$$

with τ being defined as $\min(t_{\max}, Q_{95}(S_z(t)))$, where $t_{\max}$ is the latest point in time at which an event occurred in group z and $Q_{95}(S_z(t))$ is the 95% quantile of the true survival times. Using $t_{\max}$ is necessary to ensure a fair comparison, because some estimators are only valid up to this point.[12,45]

If $\hat{S}_z(t)$ is an unbiased estimator of $S_z(t)$, $G_{\text{Bias}}(z)$ is zero, and the average of $\hat{\Delta}_{\text{Bias}}(z)$ will tend to zero as the number of simulations increases. This quantity has previously been utilized to construct hypothesis tests to formally test if two survival functions are different in a given interval.[79-81] The arithmetic mean of $\hat{\Delta}_{\text{Bias}}(z)$ over all simulation repetitions is used as an estimate of the bias overall ($\hat{G}_{\text{Bias}}(z)$). The goodness-of-fit can be similarly estimated, by utilizing the generalized mean-squared-error of the estimated survival curve and the true survival curves, defined as:[82]

$$G_{\text{MSE}}(z) = \int_0^\infty E\left(\left(S_z(t) - \hat{S}_z(t)\right)^2\right) dt = E\left(\int_0^\infty \left(S_z(t) - \hat{S}_z(t)\right)^2 dt\right). \tag{7}$$

The integral for a particular simulation run is also similarly estimated using:

$$\hat{\Delta}_{\text{MSE}}(z) = \int_0^\tau \left(S_z(t) - \hat{S}_z(t)\right)^2 dt. \tag{8}$$

The arithmetic mean of $\hat{\Delta}_{\text{MSE}}(z)$ over all simulation repetitions is used as an estimator for the real $G_{\text{MSE}}(z)$. Both $\hat{G}_{\text{Bias}}(z)$ and $\hat{G}_{\text{MSE}}(z)$ are reported with associated Monte-Carlo simulation errors, estimated as the SE of each quantity over all simulation repetitions. Additionally, the percentage of survival curves that are not monotonically decreasing as well as the percentage of estimated survival curves containing at least one point estimate falling outside of the probability bounds of 0 and 1 will be reported. For methods with existing approximate SE equations, the coverage of point-wise confidence intervals and their average width are also given.

5.4 | Results

Figure 3 shows the simulated distributions of $\hat{\Delta}_{\text{Bias}}(z = 1)$ at different sample sizes in the scenarios described above. Results for the control group ($\hat{\Delta}_{\text{Bias}}(z = 0)$) are very similar and can be seen alongside values for $\hat{G}_{\text{Bias}}(z)$ and Monte-Carlo errors in the Online Appendix. As expected, the simple Kaplan-Meier estimator is biased regardless of the sample size and scenario. When both the outcome mechanism and the treatment assignment are modeled correctly, all adjustment

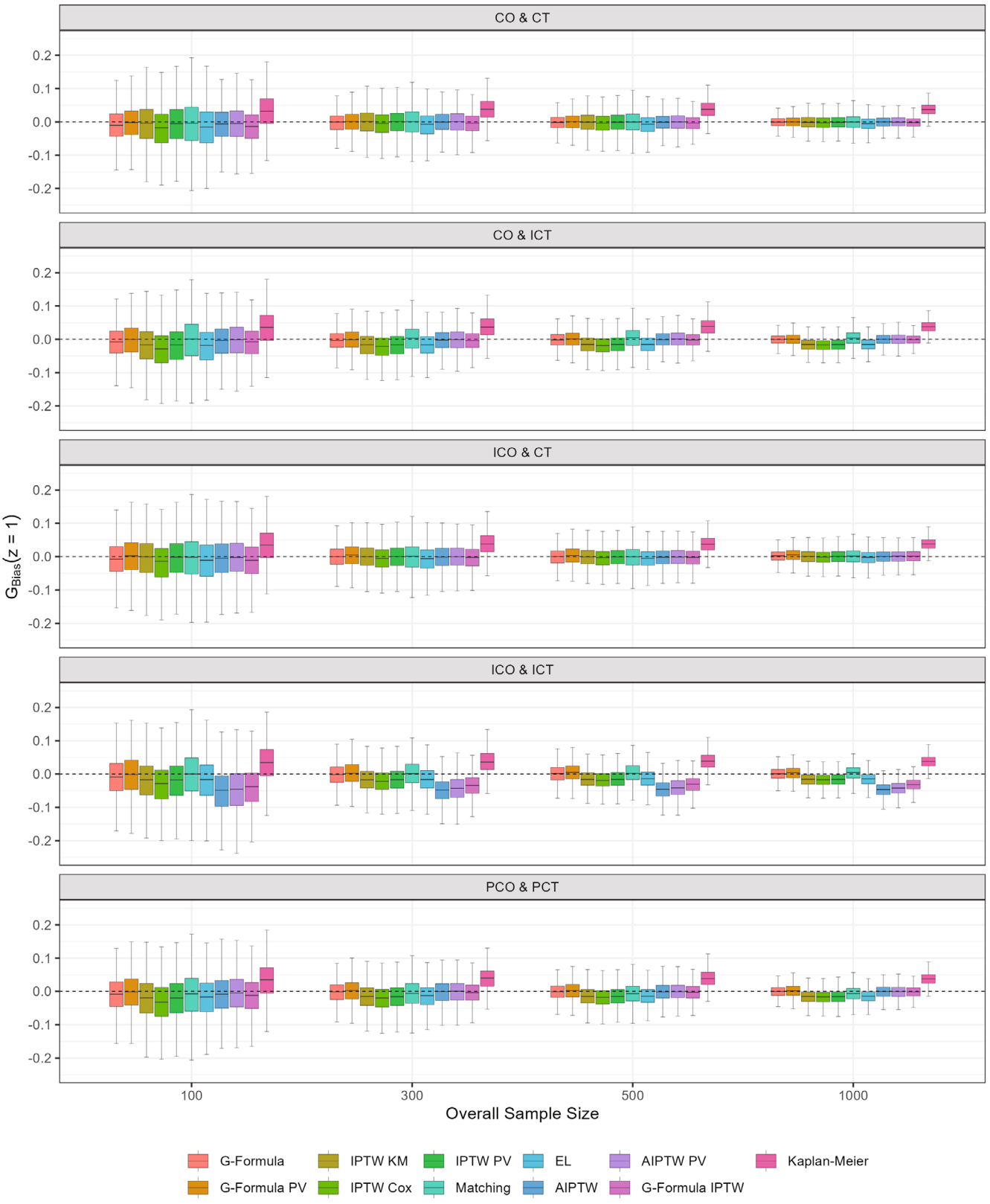

FIGURE 3 Distributions of $\hat{\Delta}_{Bias}(z = 1)$ (treatment group) for all methods in each simulation scenario with varying sample sizes. Outliers are not shown. Estimates are based on 2000 simulation repetitions.

methods show negligible amounts of bias in large samples. Methods based on Cox models, however, show systematic bias in small samples, even when the model is correctly specified. The doubly-robust property of the two AIPTW based methods can also be seen clearly from the graph. Regardless of the sample size, when at least one of the utilized models is correctly specified, the estimators remain unbiased. Interestingly, this property also extends to the PCO & PCT scenario, where the treatment-assignment model fails to include the quadratic effect of X_2 and the outcome model fails to include the quadratic effect of X_5. The current implementation of the EL method does not show this property. A small bias remains whenever the treatment-assignment model is incorrectly specified.

$\hat{\Delta}_{\text{Bias}}(z = 1)$ tends to zero for the G-Formula and the G-Formula IPTW methods when using a incorrectly specified outcome model and when both models are incorrectly specified. Therefore, these methods could be considered "unbiased" with respect to the definition of $\hat{\Delta}_{\text{Bias}}(z = 1)$ in these scenarios. This counter-intuitive result can be explained by the fact that they happen to both overestimate and underestimate the counterfactual survival probability at different points in time, cancelling out their bias. This bias can be seen more clearly in Figure 4, in which the simple average bias over

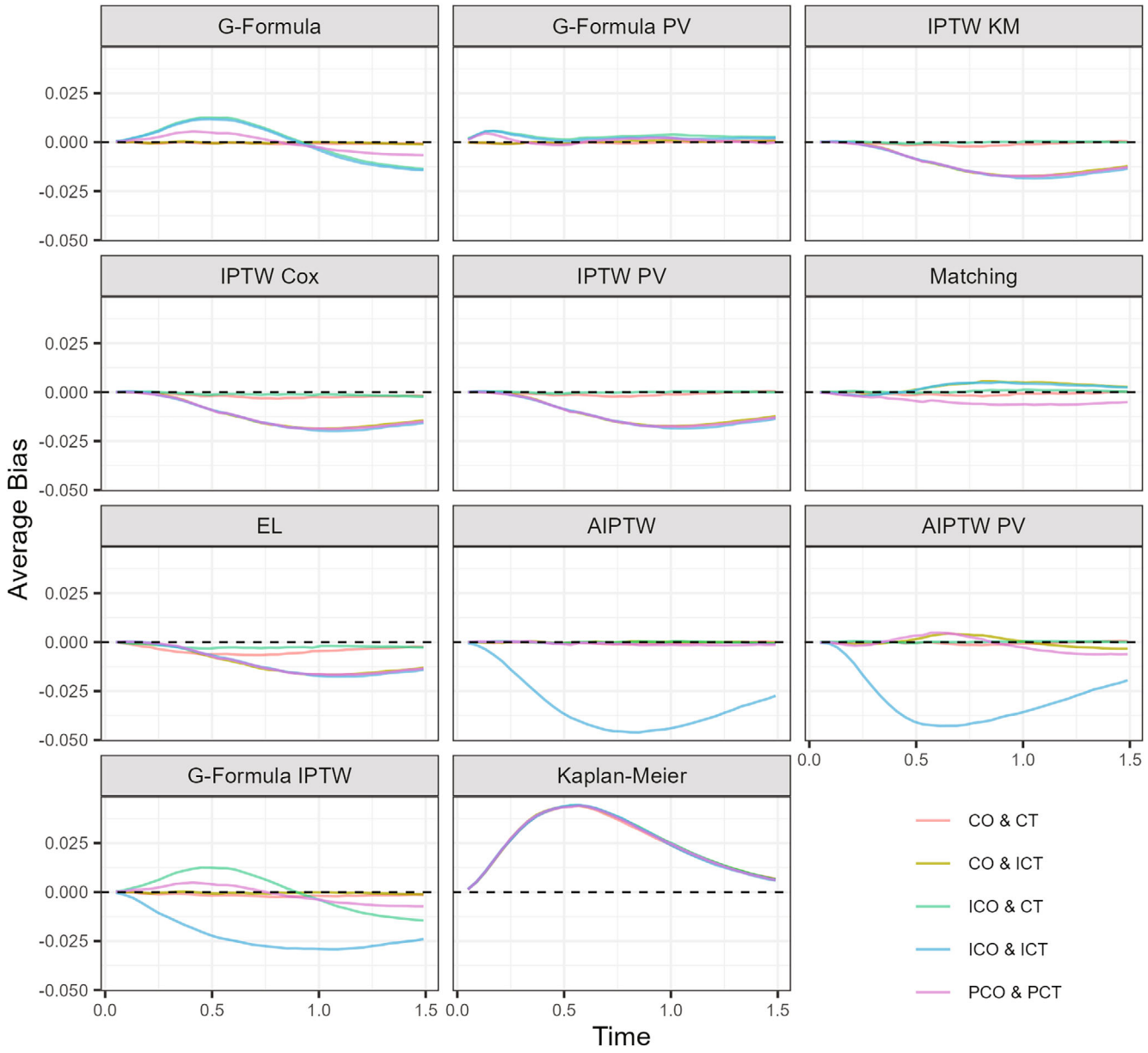

FIGURE 4 The average bias, defined as the arithmetic mean of the difference between the true survival probability and the estimated survival probability, over time for $Z = 1$ and all methods in each simulation scenario with $n = 1000$. Estimates are based on 2000 simulation repetitions.

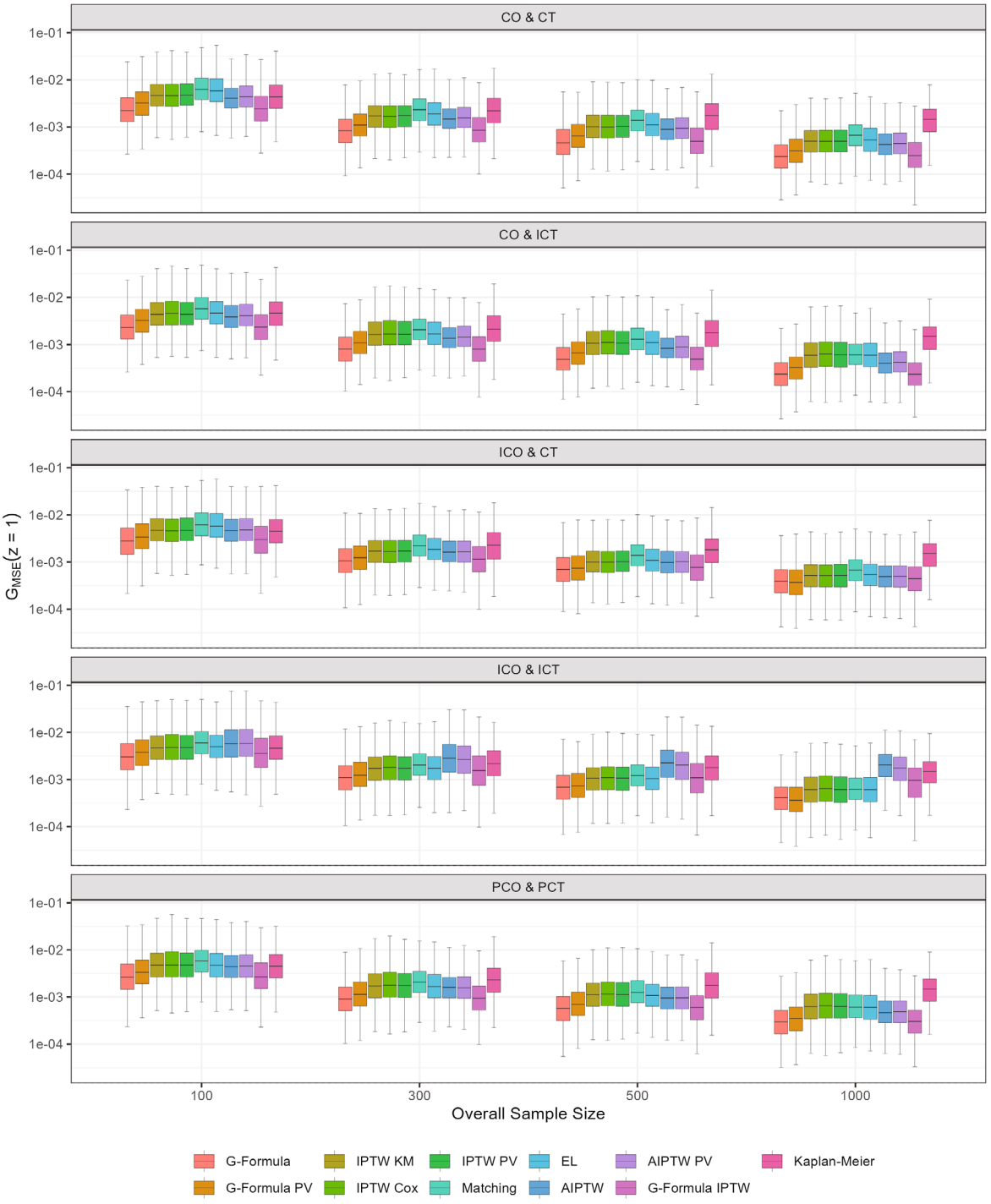

FIGURE 5 Distributions of $\hat{\Delta}_{MSE}(z = 1)$ (treatment group) on the log-scale for all methods in each simulation scenario with varying sample sizes. Outliers are not shown. Estimates are based on 2000 simulation repetitions.

time is displayed for each method and scenario for $Z = 1$ and $n = 1000$. In this figure we can also see that the G-Formula IPTW method is biased in the ICO & CT scenario as well, indicating that it does not have the doubly-robust property. All methods show systematic bias when both the outcome model and the treatment model are misspecified. More graphics displaying the bias over time can be found in the Appendix.

In Figure 5 a graph similar to Figure 3 for the goodness-of-fit in the treatment group is displayed. The same graph for the control group and values for $\hat{\Delta}_{MSE}(z)$ are shown in the Appendix. Regardless of the simulation scenario and sample size, matching shows the worst goodness-of-fit. On the opposite end of the spectrum, the G-Formula with a simple Cox

TABLE 2 Percentages of estimated survival curves in the simulation which are not strictly monotonically decreasing or which contain at least one estimate out of bounds.

		G-Formula PV		IPTW PV		AIPTW		AIPTW PV	
n	***Z***	**OOB**	**NM**	**OOB**	**NM**	**OOB**	**NM**	**OOB**	**NM**
100	0	0.00	13.10	14.00	75.75	10.35	74.20	21.85	77.00
100	1	0.00	17.35	20.90	81.35	19.70	80.95	41.10	85.05
300	0	0.00	2.90	11.85	69.40	8.25	70.60	14.45	70.55
300	1	0.00	4.40	16.60	58.40	15.55	53.55	21.65	58.40
500	0	0.00	1.95	6.50	66.30	4.70	66.65	7.65	66.75
500	1	0.00	2.65	10.90	41.20	10.95	38.80	13.15	41.90
1000	0	0.00	0.45	2.60	62.50	2.85	63.55	4.20	62.80
1000	1	0.00	0.55	2.15	24.65	2.05	23.25	2.75	23.70

Note: Based on the assumption that all models are correctly specified (Scenario: CO & CT). Using 2000 simulation repetitions.
Abbreviations: *n*, sample size; NM, Percentage of survival curves containing at least one instance of a not monotonically decreasing survival probability; OOB, Percentage of survival curves containing at least one estimate which falls outside of the 0 and 1 probability bounds; *Z*, treatment indicator.

model shows the least amount of variation, closely followed by the G-Formula using PV and a GEE model. This is true even when the Cox models are incorrectly specified. AIPTW based methods show similar amounts of variance as the IPTW based methods when only the treatment model is correctly specified, and outperform the IPTW based methods when the outcome model is also correctly specified. Empirical Likelihood Estimation has goodness-of-fit values that are very similar to the three IPTW methods. The sample size has no effect on the ranking of the methods with respect to $\hat{\Delta}_{\text{MSE}}(z)$ in any scenario.

The percentages of estimated survival curves which are not monotonically decreasing over time (NM), or which contain at least one estimate outside of the 0 and 1 probability bounds (OOB), is displayed in Table 2. The table only contains methods that are suspect to these problems (see Table 1). OOB errors occur in roughly the same frequency in the three susceptible methods. While it seems to be a very frequent problem in small sample sizes, the effect diminishes as n increases. A similar trend can be observed for the percentage of NM survival curves. The G-Formula method based on PV, however, is the least affected by this problem. Nonetheless, all four methods do exhibit some amount of these errors.

To further study when these effects occur, we plotted the percentage of OOB estimates over the entire survival curve (see Appendix). OOB errors occur much more frequently at the left and right end of the survival curve. They very rarely occur in the middle. To correct OOB estimates, the survival probability can simply be set to 1 if it is bigger than 1 and to 0 if it is negative. Isotonic regression can be used afterwards to correct NM survival curves.[67] This method uses the NM survival curve and augments it in such a way that it is changed as little as possible until it is non-decreasing, by using a weighted least squares fit subject to the monotonicity constraints. To study whether the use of these corrections has an impact on the asymptotic bias or goodness-of-fit, we applied it to all survival curves exhibiting these problems and recalculated $\hat{\Delta}_{\text{Bias}}(z)$ and $\hat{\Delta}_{\text{MSE}}(z)$. Applying these corrections resulted in a decreased bias in 53.94% of all cases and in a better goodness-of-fit in 85.88% of all cases. The changes in both bias and goodness-of-fit are, however, small on average. The average difference between the absolute value of the original $\hat{\Delta}_{\text{Bias}}(z)$ estimate and the absolute value of $\hat{\Delta}_{\text{Bias}}(z)$ after applying the corrections is $2.427934 \cdot 10^{-5}$. Similarly, the average difference between the original $\hat{\Delta}_{\text{MSE}}(z)$ and the $\hat{\Delta}_{\text{MSE}}(z)$ after applying the corrections is $8.262775 \cdot 10^{-5}$. Boxplots showing the distribution of these differences in each simulation scenario are displayed in the Appendix.

As described in Table 1, for some methods approximate equations for the SE of the causal survival probability have been derived. Where available, we used these SEs in conjunction with the normal approximation to calculate pointwise 95% confidence intervals over the whole range of the estimated survival curves. Figure 6 shows the percentage of these confidence intervals containing the true survival probability over time for different sample sizes and methods in the CO & CT scenario. Regardless of method and sample size, the coverage is below 95% for very early points in time and sometimes too high at the end of the curves. In most parts of the curves however, the confidence interval coverage is appropriate at every sample size and for every method. A similar graphic for the confidence interval width is included in the Online Appendix.

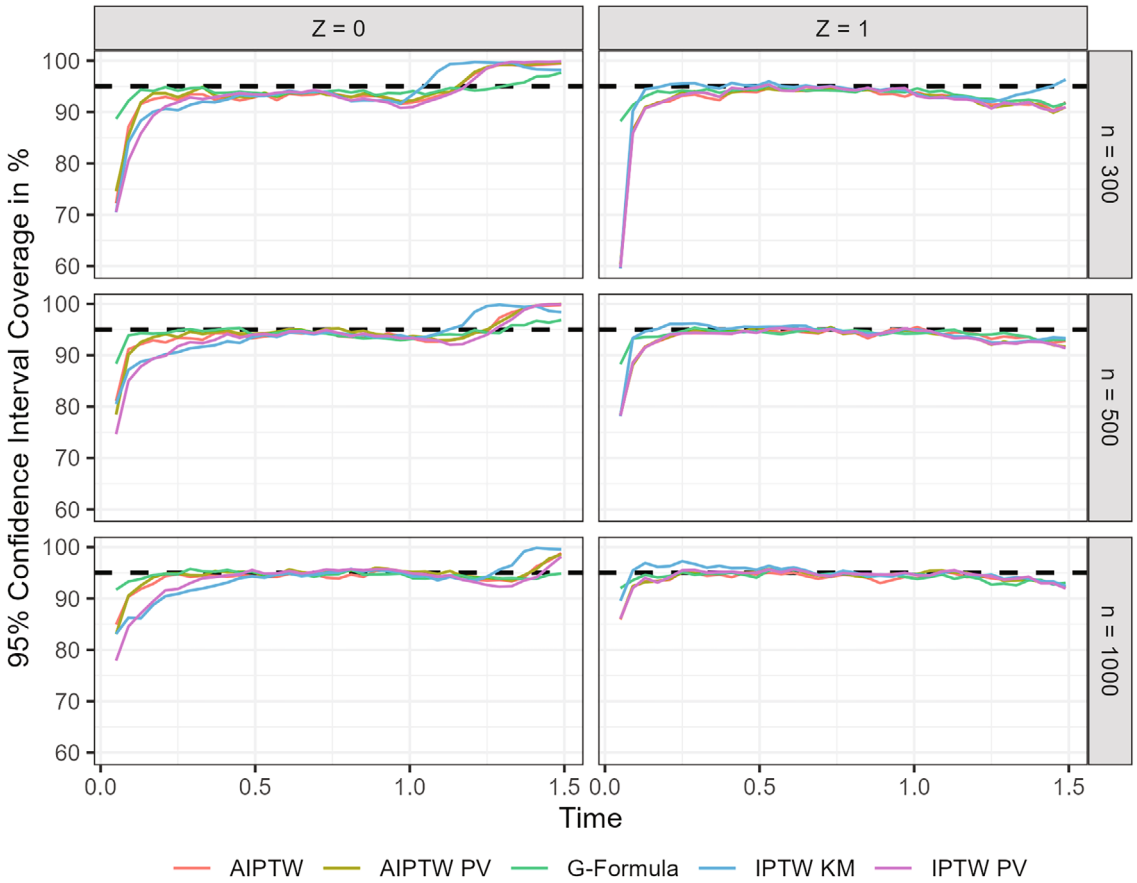

FIGURE 6 The percentage of 95% confidence intervals which contain the true survival probability over time for each method in the CO & CT simulation scenario for both treatment groups and varying sample sizes. Estimates are based on 2000 simulation repetitions.

6 | DISCUSSION

In this article, we compared different methods for calculating confounder-adjusted survival curves with right-censored data under the assumption of proportional hazards. Our main focus was on the bias and the goodness-of-fit in different scenarios. The simulations showed that when used properly, all considered methods produce unbiased estimates for the whole survival curve in medium to large sample sizes. Methods utilizing the outcome mechanism were shown to be more efficient than methods relying on the treatment mechanism only. AIPTW based methods showed amounts of variation that are comparable to IPTW based methods when only one model was correctly specified and slightly outperformed the IPTW methods when both models were correct. These results are consistent with the previous literature on this topic.[19,20] However, in contrast to the study by Chatton et al,[16] we found a small amount of bias when using the G-Formula IPTW method when a confounder was incorrectly modeled the outcome model. It follows that this method does not have the doubly-robust property when estimating the counterfactual survival curve. A possible reason for this may be that the addition of the weights results in a violation of the proportional hazards assumption. Using a different outcome model might therefore give better results.

We were also able to show systematic bias in methods relying on a Cox model in small sample sizes ($n < 300$), even when it was correctly specified. An exception to this rule is the AIPTW method, which stays unbiased in small sample sizes when either the Cox model or the treatment assignment model are correctly specified. There were only small differences between the performance of PV based methods and their non-PV based counterparts. While AIPTW and PV based methods showed significant problems with monotonicity and estimates falling out of the 0 and 1 probability bounds, they should not be disregarded. Simple corrections, including isotonic regression and truncation, can be applied without introducing bias or loosing efficiency.[67]

This study has several limitations. First, we only considered situations in which the true time-to-event process can be described using a Cox model, without time-varying confounders. This is an optimistic depiction of reality at best, but

it reflects the assumptions made by most researchers in practice. Furthermore, we only considered a binary treatment variable, even though in reality, multi-arm studies are very common. We did this for the case of simplicity and because not all considered methods currently support categorical treatment variables.[14] Additionally, the generalized bias we used to judge the performance of the methods clearly has some shortcomings. If a method both overestimates and underestimates the true counterfactual survival probability in different points in time, the method can theoretically have a generalized bias of zero, despite it being biased at almost every point in time. We substituted the statistics with plots of the time-specific bias over time to eliminate this problem.

Nevertheless, we were able to show that in the considered scenarios, all methods consistently outperformed the naive Kaplan-Meier estimates. We therefore think that the methods discussed here should be used instead of the standard Kaplan-Meier estimator when analysing observational data. Based on the results of this study and previous discussion on the topic[56,58] we recommend using AIPTW based methods, because they possess the doubly-robust property and showed goodness-of-fit similar to IPTW based methods. Although the G-Formula and G-Formula PV methods showed better goodness-of-fit overall, they do rely on one correctly specified model. When using the Cox model, this entails including all independent predictor variables in addition to the confounders, which can be difficult in practice. A drawback of the AIPTW and AIPTW PV methods is that they are more complex to use than other methods, because an implementation of the method itself and of isotonic regression is required. We believe, however, that this problem is mitigated by the user-friendly **R**-Code implementations in the `riskRegression`[44] and `adjustedCurves`[77] packages.

Although the EL method is also doubly-robust in theory, the only currently available implementation does not have this property, as demonstrated in our study. If such an implementation becomes available in the future, it would be a viable alternative to the AIPTW based methods because it does not rely on any kind of model. When there are known unmeasured confounders, the methods discussed here are insufficient to obtain unbiased estimates. Instrumental variable based methods might be preferable in this case.[28]

ACKNOWLEDGEMENTS

We would like to thank Xiaofei Wang and Fangfang Bai for supplying us with their R-Code used for empirical likelihood estimation. We also want to thank Jixian Wang for supplying us with the R-Code used for the calculation of the augmented inverse probability of treatment weighted estimator based on pseudo-values. Additionally we want to thank the anonymous reviewers for their extensive input, which greatly improved this manuscript. Last but not least, we want to thank Karl-Heinz Jöckel, Jale Basten, Marianne Tokic and Henrik Rudolf for their helpful comments and suggestions. Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

R-code to perform and analyse the Monte-Carlo simulations presented in this article is available as part of the online version of this article. R-code implementing the presented methods is publicly available in our `adjustedCurves` package on CRAN (https://cloud.r-project.org/web/packages/adjustedCurves/index.html). The data used in the illustrative example of this study are available on suitable request from the getABI study group. The data are not publicly available due to German data privacy laws.

ORCID

Robin Denz https://orcid.org/0000-0002-2682-5268

REFERENCES

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
2. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol*. 1974;66(5):688-701.
3. Rubin DB. Randomization analysis of experimental data: the fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591-593.
4. Cox DR. Regression models and life-tables. *J R Stat Soc B Methodol*. 1972;34(2):187-220.
5. Poole C. On the origin of risk relativism. *Epidemiology*. 2010;21(1):3-9.

6. Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010;21(1):13-15.
7. De Neve J, Gerds TA. On the Intrepretation of the hazard ratio in Cox regression. *Biom J*. 2020;62:742-750.
8. Davis CR, McNair AGK, Brigic A, et al. Optimising methods for communicating survival data to patients undergoing cancer surgery. *Eur J Cancer*. 2010;46:3192-3199.
9. Zipkin DA, Umscheid CA, Keating NL, et al. Evidence-based risk communication: a systematic review. *Ann Intern Med*. 2014;161(4):270-280.
10. Dey T, Mukherjee A, Chakraborty S. A practical overview and reporting strategies for statistical analysis of survival studies. *Chest*. 2020;158(1, Supplement):S39-S48.
11. Makuch RW. Adjusted survival curve estimation using covariates. *J Chronic Dis*. 1982;35(6):437-443.
12. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Comput Methods Programs Biomed*. 2004;2003(75): 45-49.
13. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med*. 2014;33:1242-1258.
14. Wang X, Bai F, Pang H, George SL. Bias-adjusted Kaplan-Meier survival curves for marginal treatment effect in observational studies. *J Biopharm Stat*. 2019;29(4):592-605.
15. Zhang M, Schaubel DE. Contrasting treatment-specific survival using double-robust estimators. *Stat Med*. 2012;31(30):4255-4268.
16. Chatton A, Borgne FL, Leyrat C, Foucher Y. G-computation and doubly robust standardisation for continuous-time data: a comparison with inverse probability weighting. *Stat Methods Med Res*. 2021;31(4):706-718.
17. Andersen PK, Syriopoulou E, Parner ET. Causal inference in survival analysis using pseudo-observations. *Stat Med*. 2017;36: 2669-2681.
18. Wang J. A simple, doubly robust, efficient estimator for survival functions using pseudo observations. *Pharm Stat*. 2018;17:38-48.
19. Cai W, van der Laan MJ. One-step targeted maximum likelihood estimation for time-to-event outcomes. *Biometrics*. 2020;76: 722-733.
20. Chatton A, Borgne FL, Leyrat C, Foucher Y. G-computation and inverse probability weighting for time-to-event outcomes: a comparative study. arXiv preprint arXiv:2006.16859v1, 2020.
21. Neyman JS. Statistical problems in agricultural experiments. *J R Stat Soc B*. 1923;2:107-180.
22. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1): 41-55.
23. Neugebauer R, van der Laan MJ. Why prefer double robust estimators in causal inference. *J Stat Plan Inference*. 2005;129:405-426.
24. Guo S, Fraser MW. *Propensity Score Analysis: Statistical Methods and Applications*. 2nd ed. Los Angeles, CA: Sage; 2015.
25. Jiang H, Symanowski J, Qu Y, Ni X, Wang Y. Covariate-adjusted non-parametric survival curve estimation. *Stat Med*. 2010;30: 1243-1253.
26. Zeng D. Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Ann Stat*. 2004;32(4):1533-1555.
27. Clare PJ, Dobbins TA, Mattick RP. Causal models adjusting for time-varying confounding: a systematic review of the literature. *Int J Epidemiol*. 2019;48(1):254-265.
28. Martínez-Camblor P, MacKenzie TA, Staiger DO, Goodney PP, O'Malley AJ. Summarizing causal differences in survival curves in the presence of unmeasured confounding. *Int J Biostat*. 2020;17(2):223-240.
29. Chang IM, Gelman R, Pagano M. Corrected group prognostic curves and summary statistics. *J Chronic Dis*. 1982;35:669-674.
30. Nieto FJ, Coresh J. Adjusting survival curves for confounders: a review and a new method. *Am J Epidemiol*. 1996;143(10): 1059-1068.
31. Ghali WA, Quan H, Brant R, et al. Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models. *JAMA*. 2001;286(12):1494-1497.
32. Cupples LA, Gragnon DR, Ramaswamy R, D'Agostino R. Age-adjusted survival curves with application in the Framingham study. *Stat Med*. 1995;14:1731-1744.
33. Gregory WM. Adjusting survival curves for imbalances in prognostic factors. *Br J Cancer*. 1988;58:202-204.
34. Amato DA. A generalized Kaplan-Meier estimator for heterogenous populations. *Commun Stat Theory Methods*. 1988;17(1): 263-286.
35. Moore KL, van der Laan MJ. Application of time-to-event methods in the assessment of safety in clinical trials. In: Peace KE, ed. *Design and Analysis of Clinical Trials with Time-to-Event Endpoints*. Boca Raton, FL: CRC Press; 2009.
36. Stitelman OM, van der Laan MJ. Collaborative targeted maximum likelihood for time to event data. *Int J Biostat*. 2010;6(1):21.
37. Guerra SF, Schnitzer ME, Forget A, Blais L. Impact of discretization of the timeline for longitudinal causal inference methods. *Stat Med*. 2020;39(27):4069-4085.
38. Sofrygin O, Zhu Z, Schmittdiel JA, et al. Targeted learning with daily EHR data. *Stat Med*. 2019;38:3073-3090.
39. Westling T, Luedtke A, Gilbert P, Carone M. Inference for treatment-specific survival curves using machine learning. arXiv preprint arXiv:2106.06602v1, 2021.
40. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat*. 1978;6(1):34-58.
41. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Math Model*. 1986;7:1393-1512.
42. Breslow N. Discussion of the paper by D.R. Cox. *J R Stat Soc B*. 1972;34(2):216-217.

43. Zhang X, Loberiza FR, Klein JP, Zhang MJ. A SAS macro for estimation of direct adjusted survival curves based on a stratified Cox regression model. *Comput Methods Programs Biomed*. 2007;88:95-101.
44. Ozenne BMH, Scheike TH, Stærk L. On the estimation of average treatment effects with right-censored time to event outcome and competing risks. *Biom J*. 2020;62:751-763.
45. Xie J, Liu C. Adjusted Kaplan-Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Stat Med*. 2005;24:3089-3110.
46. McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol Methods*. 2004;9(4):403-425.
47. Imai K, Ratkovic M. Covariate balancing propensity score. *J R Stat Soc B*. 2014;76(1):243-263.
48. Chatton A, Borgne FL, Leyrat C, et al. G-computation, propensity score-based methods, and targeted maximum likelihood estimator for causal inference with different covariates sets: a comparative simulation study. *Sci Rep*. 2020;10:9219.
49. Raad H, Cornelius V, Chan S, Williamson E, Cro S. An evaluation of inverse probability weighting using the propensity score for baseline covariate adjustment in smaller population randomised controlled trials with a continuous outcome. *BMC Med Res Methodol*. 2020;20:70.
50. Winnett A, Sasieni P. Adjusted Nelson-Aalen estimates with retrospective matching. *J Am Stat Assoc*. 2002;97(457):245-256.
51. Galimberti S, Sasieni P, Valsecchi MG. A weighted Kaplan-Meier estimator for matched data with application to the comparison of chemotherapy and bone-marrow transplant in leukaemia. *Stat Med*. 2002;21:3847-3864.
52. Austin PC, Thomas N, Rubin DB. Covariate-adjusted survival analyses in propensity-score matched samples: imputing potential time-to-event outcomes. *Stat Methods Med Res*. 2020;29(3):728-751.
53. Sekhon JS. Multivariate and propensity score matching software with automated balance optimization: the matching package for R. *J Stat Softw*. 2011;42(7):1-52.
54. Borgne FL, Giraudeau B, Querard AH, Giral M, Foucher Y. Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. *Stat Med*. 2016;35:1103-1116.
55. Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Stat Methods Med Res*. 2016;25(5):2214-2237.
56. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell NP, Dietz K, Farewell VT, eds. *AIDS Epidemiology: Methodological Issues*. New York: Springer Science + Business Media; 1992:297-331.
57. Hubbard AE, van der Laan MJ, Robins JM. Nonparametric locally efficient estimation of the treatment specific survival distribution with right censored data and covariates in observational studies. In: Halloran ME, Berry D, eds. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. New York: Springer Science + Business Media; 2000:135-177.
58. Bai X, Tsiatis AA, O'Brien SM. Doubly-robust estimators of treatment-specific survival distributions in observational studies with stratified sampling. *Biometrics*. 2013;69:830-839.
59. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci*. 2007;22(4):523-539.
60. Vansteelandt S, Keiding N. Invited commentary: G-computation-lost in translation? *Am J Epidemiol*. 2011;173(7):739-742.
61. Lee D, Yang S, Wang X. Generalizable survival analysis of randomized controlled trials with observational data. arXiv preprint arXiv:2201.06595v1, 2022.
62. Owen AB. *Empirical Likelihood*. Boca Raton, FL: CRC Press; 2001.
63. Andersen PK, Perme MP. Pseudo-observations in survival analysis. *Stat Methods Med Res*. 2010;19:71-99.
64. Zeng S, Li F, Hu L, Li F. Propensity score weighting analysis of survival outcomes using pseudo-observations. arXiv preprint arXiv:2103.00605v1, 2021.
65. Klein JP, Gerster M, Andersen PK, Tarima S, Perme MP. SAS and R functions to compute pseudo-values for censored data regression. *Comput Methods Programs Biomed*. 2008;89(3):289-300.
66. Overgaard N, Parner ET, Pedersen J. Asymptotic theory of generalized estimating equations based on Jack-knife pseudo-observations. *Ann Stat*. 2017;45(5):1988-2015.
67. Westling T, van der Laan MJ, Carone M. Correcting an estimator of a multivariate monotone function with isotonic regression. *Electron J Stat*. 2020;14:3032-3069.
68. Lee PN, Forey BA, Coombs KJ. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC Cancer*. 2012;12:385.
69. Carter BD, Abnet CC, Feskanich D, et al. Smoking and mortality: beyond established causes. *N Engl J Med*. 2015;372(7):631-640.
70. Pearl J, Mackenzie D. *The Book of why: the New Science of Cause and Effect*. London: Penguin Books; 2018.
71. Li K, Yao C, Di X, et al. Smoking and risk of all-cause deaths in younger and older adults. *Medicine*. 2016;95(3):e2438.
72. Gutterman S. Mortality of smoking by gender. *N Am Actuar J*. 2015;19(3):200-223.
73. Schnohr C, Højbjerre L, Riegels M, et al. Does educational level influence the effects of smoking, alcohol, physical activity, and obesity on mortality? A prospective population study. *Scand J Public Health*. 2004;32(4):250-256.
74. getABI Study Group. getABI: German epidemiological trial on ankle brachial index for elderly patients in family practice to Dedect peripheral arterial disease, significant marker for high mortality. *Vasa*. 2002;31(4):241-248.
75. Diehm C, Schuster A, Allenberg JR, et al. High prevalence of peripheral arterial disease and co-morbidity in 6880 primary care patients: cross-sectional study. *Atherosclerosis*. 2004;172(1):95-105.
76. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Stat Med*. 2005;24(11):1713-1723.

77. Denz R. adjustedCurves: confounder-adjusted survival curves and cumulative incidence functions. R package version 0.9.0; 2022. https://cran.r-project.org/package=adjustedCurves.
78. Martinussen T, Vansteelandt S. On collapsibility and confounding bias in Cox and Aalen regression models. *Lifetime Data Anal*. 2013;19:279-296.
79. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*. 1989;45(2):497-507.
80. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: large sample and optimality considerations. *J R Stat Soc B*. 1991;53(2):341-352.
81. Zhao L, Tian L, Uno H, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring and analyzing a comparative clinical study. *Clin Trials*. 2012;9(5):570-577.
82. Klein JP, Moeschberger ML. The robustness of several estimators of the survivorship function with randomly censored data. *Commun Stat Simul Comput*. 1989;18(3):1087-1112.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Denz R, Klaaßen-Mielke R, Timmesfeld N. A comparison of different methods to adjust survival curves for confounders. *Statistics in Medicine*. 2023;42(10):1461-1479. doi: 10.1002/sim.9681