Biometrical Journal

Analysis of survival data with nonproportional hazards: A comparison of propensity-score-weighted methods

Elizabeth A. Handorf 1 | Marc C. Smaldone 2 | Sujana Movva 3 | Nandita Mitra 4

Abstract

One of the most common ways researchers compare cancer survival outcomes across treatments from observational data is using Cox regression. This model depends on its underlying assumption of proportional hazards, but in some real-world cases, such as when comparing different classes of cancer therapies, substantial violations may occur. In this situation, researchers have several alternative methods to choose from, including Cox models with time-varying hazard ratios; parametric accelerated failure time models; Kaplan–Meier curves; and pseudo-observations. It is unclear which of these models are likely to perform best in practice. To fill this gap in the literature, we perform a neutral comparison study of candidate approaches. We examine clinically meaningful outcome measures that can be computed and directly compared across each method, namely, survival probability at time $T$, median survival, and restricted mean survival. To adjust for differences between treatment groups, we use inverse probability of treatment weighting based on the propensity score. We conduct simulation studies under a range of scenarios, and determine the biases, coverages, and standard errors of the average treatment effects for each method. We then demonstrate the use of these approaches using two published observational studies of survival after cancer treatment. The first examines chemotherapy in sarcoma, which has a late treatment effect (i.e., similar survival initially, but after 2 years the chemotherapy group shows a benefit). The other study is a comparison of surgical techniques for kidney cancer, where survival differences are attenuated over time.

KEYWORDS

inverse probability of treatment weighting, nonproportional hazards, observational studies

1 INTRODUCTION

Many cancer outcomes studies use observational databases, such as SEER-Medicare or the National Cancer Database (Enewold et al., 2020; Boffa et al., 2017), to compare survival outcomes between different treatment or exposure groups. A common approach to model survival data with censoring and potentially confounding covariates is the Cox proportional hazards regression model (Cox, 1972). However, the standard Cox model makes a strong assumption that the hazards are proportional between groups over time (a constant hazard ratio [HR]) (Klein & Moeschberger, 2003). This assumption
may not hold in real-world studies. In recent years, this issue has arisen more frequently in the oncology literature, as nonproportionality is a common feature in studies of immunotherapies (Jiménez, 2022; Rahman et al., 2019).

There are a variety of models for censored outcomes that do not assume proportional hazards. Nonparametric methods, including the ubiquitous Kaplan–Meier method, do not make assumptions about the functional form of the survival curves, and therefore, accommodate nonproportionality (Collett, 2015; Kaplan & Meier, 1958). A recent method that uses pseudo-observations can also be used without parametric assumptions. This method, using a missing data framework, uses a consistent estimator to replace each observation with a pseudo-observation (i.e., its estimate when it is excluded from the sample) (Andersen et al., 2004; Andersen & Pohar Perme, 2010; Andersen et al., 2017; Ambrogiet al., 2022; Sachs & Gabriel, 2022). There are also parametric and semiparametric approaches that accommodate nonproportionality. The Cox model can be modified to include time-dependent hazards, relaxing the assumption of proportionality (Collett, 2015). Accelerated failure time (AFT) models are a different class of regression methods for censored data, which assume a linear model for log-time. They can be modeled fully parametrically, and for many choices of error distributions, the models are nonproportional (Collett, 2015).

Given the number of choices, a natural question is which method performs best in real-world conditions, particularly when confounding is present. As standard implementations of nonparametric approaches do not readily incorporate covariates, in this paper, we focus on approaches that can use propensity score weighting to improve balance and address confounding (including parametric AFT models, Cox models, pseudo-observations, and Kaplan–Meier curves). The propensity score is commonly used for observational studies and has many advantages as we discuss in Section 2. We then describe several methods for analysis of survival data that do not rely on proportional hazards. We focus on clinically relevant estimands and compare the performance of several parametric, semiparametric, and nonparametric approaches. We focus on methods that can (1) be used with inverse probability of treatment weighting (IPTW), (2) identify the estimands of interest, and (3) are readily implemented by applied researchers using currently available software. In this paper, we use realistic simulated scenarios in a neutral comparison study of possible methods. Our results provide insight on the performance of these methods in clinical applications.

This work is motivated by two recently published clinical studies in which the treatment effect exhibited nonproportional hazards. In a study of chemotherapy for soft-tissue sarcoma, there was a negligible survival benefit until about 2 years after the start of treatment (Movva et al., 2015). In another study comparing two different surgical techniques in early-stage renal cancer, one approach had a strong benefit immediately after surgery, but the benefit decreased over time (Ristau et al., 2018). These two forms of deviation from proportionality served as the motivation for this study. The data for both of these studies came from an observational source, the National Cancer Database (Boffa et al., 2017). Therefore, both were likely subject to confounding by indication; that is, the choice of treatment was informed by factors that are also associated with survival.

2 PROPENSITY SCORE WEIGHTING

The propensity score, first proposed by Rosenbaum and Rubin (1983, 1984) is defined as the probability of receiving treatment given a set of covariates,

\[ e = P(Z = 1 | X), \]

where \( e \) is the propensity score, \( Z \) is a binary indicator for treatment, and \( X \) is a vector of (measured) covariates.

Rosenbaum and Rubin showed that the propensity score is a balancing score, that is, conditional on the propensity score, the treatment is independent of the covariates. This has important implications for control of confounding in observational studies; namely, if the propensity score is defined based on all confounding covariates, controlling for the propensity score will control for all sources of confounding.

Propensity score methods have a number of benefits over traditional regression adjustment. First, they provide an easy way to check for sufficient covariate overlap and balance between the treated and control groups (Austin, 2011). Second, one can use flexible and modern approaches, such as ensemble learning approaches and other machine learning models to estimate the propensity scores (Cannas & Arpino, 2019; Pirracchio et al., 2015; van der Laan et al., 2007; Zou et al., 2021). This can make the estimates less vulnerable to bias from an incorrect functional form. Propensity scores also allow for the use of doubly robust methods (Funk et al., 2011). Third, propensity score weighting approaches provide marginal, not conditional, estimates, so the interpretation of a treatment effect does not depend on a subject's covariate values.
Marginal estimates are therefore more interpretable and useful to clinicians and policymakers (Austin & Stuart, 2015, 2017; Lunceford & Davidian, 2004; Mao et al., 2018). There are drawbacks to using propensity score methods as well. First, and most importantly, propensity score methods cannot overcome biases from unmeasured confounders. Further, they may be less familiar and less readily understood by clinical readership. The propensity score can also be used in several different ways (e.g., stratification, matching, and weighting), and changing the method may change the estimand, producing results with different interpretations. Propensity score methods therefore must be conducted by analysts with expertise in the area, and clearly described when publishing results (Deb et al., 2016).

There are several ways to use propensity scores in analyses of survival data. Here, we focus on weighting, specifically IPTW. IPTW creates a synthetic sample where covariates are unrelated to treatment assignment, so that treatment effects in the weighted samples can be summarized and compared directly.

To implement an IPTW analysis, one must first obtain the estimated propensity scores for each subject ($\hat{e}_i$). This can be accomplished via a number of approaches, including logistic regression, random forests, or ensemble methods. The IPTW for subject $i$ is then defined as

$$\hat{w}_i = \frac{Z_i}{\hat{e}_i} + \frac{1 - Z_i}{1 - \hat{e}_i}.$$ 

As $Z_i$ is binary, if subject $i$ is treated, their weight is one over their probability of being treated (based on their covariates). If the subject did not receive the treatment, their weight is one over their probability of receiving the control. Therefore, treated subjects who were less likely to be treated have larger weights, as do control subjects who were less likely to receive the control, thus creating a balanced pseudo-population. In subsequent analyses, these weights are used to estimate the expected effects under the treatment and control conditions. It is important to note that one should check for covariate balance between treatment arms in the pseudo-population. Often, a standardized mean difference of 0.1 or less between arms is taken to indicate good balance (Austin, 2011).

### 3 | SURVIVAL METHODS UNDER NONPROPORTIONAL HAZARDS

#### 3.1 | Quantifying differences between groups with nonproportionality

In the setting of survival outcomes with nonproportionality, we must first choose an estimand. It should be readily interpretable and meaningful given the observed data features. If we fit a proportional hazards model to nonproportional data, we would obtain a single HR, but this is actually a weighted average of the HRs at each observed failure time. The weighted average is problematic because the weights depend on the censoring distribution. This means that factors unrelated to the treatment effect, such as the accrual times and length of follow-up, will affect the estimated HR when hazards are nonproportional (Horiguchi et al., 2019; Kalbfleisch & Prentice, 1981; Uno et al., 2014). Alternative estimands include the average HR, an average weighted by the number of patients at risk (Schemper et al., 2009).

The difficulties inherent in reporting HRs in the setting of nonproportionality can be avoided by focusing on measures that can be defined directly from the survival functions for the treated and control arms, $S_1(t)$ and $S_0(t)$, respectively.

The simplest approach compares differences in survival probabilities at a particular time $t = T$.

$$\Delta S_T = S_1(t = T) - S_0(t = T).$$

Quantiles of the survival functions can also be compared. For instance, if the survival curves are defined for $S(t) = 0.5$, median survival times can be compared.

$$\Delta T_{50} = (T : S_1(t = T) = 0.5) - (T : S_0(t = T) = 0.5).$$

These measures are readily understood by a clinical audience, but they do not comprehensively measure survival effects over the length of follow-up. Another measure is restricted mean survival (RMS), the mean survival up through some time $T_{max}$ where $S(t)$ is defined (Conner et al., 2019). Often, $T_{max}$ can be chosen as the largest observed time, although this does rely on the assumption that the censoring distribution meets certain (mild) conditions (Tian et al., 2020). In this context, one should choose $T_{max}$ to be no more than the largest time that is observed in both arms. The RMS for each arm can be
found using the area under the survival curve, and the difference in RMS between the treated and control conditions is

\[ \Delta \text{RMS}_{T_{\text{max}}} = \int_0^{T_{\text{max}}} S_1(t) dt - \int_0^{T_{\text{max}}} S_0(t) dt. \]

RMSs (and related extensions) have become increasingly popular. There are several specialized methods for estimating RMS, including doubly robust and flexible parametric models (Royston & Parmar, 2011; Zhang & Schaubel, 2012).

Each of the measures described above are always interpretable, regardless of whether proportionality holds. In the subsequent sections, we focus on methods that can (1) accommodate IPTW, (2) allow the estimation of the measure of interest (\( \Delta S_T, \Delta T_{50}, \) and \( \Delta \text{RMS}_{T_{\text{max}}} \)), and (3) are readily implemented using published software. We note that some useful methods, such as those that only support estimation of RMS, are not included in the paper, as they do not meet all the inclusion criteria listed above.

### 3.2 Cox model

Although the Cox model uses the assumption of proportionality, it is often robust to small deviations from this assumption. However, if the sample size is large, as is common in registry studies, small deviations from proportionality may be statistically significant when assessed using standard diagnostic tests (e.g., Cox–Snell residuals, Schoenfeld residuals) (Grambsch & Therneau, 1994). Therefore, the Cox model may still be useful in some cases where deviations from proportionality are demonstrated. The Cox model is defined as follows:

\[ h_i(t \mid Z_i) = h_0(t) \exp(\beta Z_i). \]

IPTW can be readily accommodated in the Cox model, by weighting the contribution of each observation to the partial likelihood function:

\[ L(\beta) = \prod_{i=1}^{N} \left( \frac{\exp(\beta Z_i)}{\sum_{j \in R(t_i)} \hat{w}_j \exp(\hat{\beta} Z_j)} \right) \hat{w}_i, \]

where \( R(t_i) \) is the risk set at time \( t_i \). The parameter \( \hat{\beta} \) is the average of the log-HRs at each failure time. Importantly, one can find \( \hat{S}(t) \) based on the estimates \( \hat{h}_0(t) \) and \( \hat{\beta} \).

We can also extend the standard Cox model, relaxing the assumption of proportionality, by allowing the HR to vary as a function of time.

\[ h_i(t \mid Z_i) = h_0(t) \exp(\beta Z_i + f(Z_i, t)). \]

In our simulations, we consider two particular cases: (1) allowing the hazard to vary by log-time, where the parameter \( \kappa \) defines the strength of the nonproportionality of the treatment effect.

\[ f(Z_i, t) = \kappa Z_i \log(t), \]

and (2) using a piecewise constant (PWC) treatment effect

\[ f(Z_i, t) = \begin{cases} 0 & 0 \leq t < C_1 \\ \kappa_1 Z_i & C_1 \leq t < C_2 \\ \kappa_2 Z_i & C_2 \leq t, \end{cases} \]

where \( C_1 \) and \( C_2 \) are the prespecified times at which the constant hazards are allowed to change. These models can all be implemented with the `survival` package in R (Therneau, 2020).
3.3 Parametric AFT models

AFT models assume a linear model for $\log(t)$ (Buckley & James, 1979; Collett, 2015; Jin et al., 2006; Wei, 1992).

$$\log(T_i) = \mu + \beta Z_i + \sigma \varepsilon_i,$$

where $T_i$ is the realization of time $t$ for subject $i$. Here, we consider parametric models, where the error term $\varepsilon$ is assumed to take on a given distribution. A common choice is the Gumbel distribution, which yields a Weibull hazard model; however, the Weibull model is also a proportional hazards model when the shape parameter is constant. We can accommodate nonproportionality by allowing the HR to change over time.

$$h_i(t | Z) = \exp[-(\beta Z_i + f(Z_i, t))] \lambda t^{\gamma - 1},$$

where $\lambda$ is the scale and $\gamma$ is the shape parameter. If we let

$$f(Z_i, t) = \nu Z_i \log(t),$$

then (1) can be rewritten as

$$\log(T_i) = \frac{1}{\gamma} + \frac{\nu Z_i}{\gamma} (-\log(\lambda) + \beta Z_i + \varepsilon_i).$$

$$\varepsilon_i \sim \text{Gumbel}$$

If we allow the shape to vary by treatment, this allows the HR to change over log-time.

Another flexible alternative is the generalized gamma model. Marshall and Olkin (1997), a three-parameter gamma model. In the AFT model specification of Cox et al. (2007), $(T_i)$ has a generalized gamma distribution with the parameters $(\mu + \beta Z), \sigma (\sigma > 0)$, and $\lambda \in \mathbb{R}$, such that

$$\log(T_i) = \mu + \beta Z_i + \sigma \varepsilon_i,$$

$$\varepsilon_i \sim \left(\frac{1}{\varepsilon_i \Gamma(\lambda^{-2})}\right) (\lambda^{-2} \varepsilon_i^{\lambda-1})^{\lambda-2} \exp(-\lambda^{-2} \varepsilon_i^{\lambda}).$$

This model encompasses other distributions as special cases; when $\lambda = 1$, we have a Weibull AFT model, if $\lambda = 0$, it results in a log-normal model, and when $\lambda = \sigma$, we have the standard Gamma model.

For these AFT models, the weights $\hat{w}_i$ can be applied during estimation of parameters as frequency weights (i.e., the weights that create the pseudo-population). Each subjects’ weight therefore determines their contribution to the likelihood function. Here, we focus on the three-parameter Weibull and generalized gamma AFT models, but a wide range of AFT models can be fit in R using the flexsurv package. Other parametric models, including spline-based models such as the Royston–Parmar model, can also be fit using this package (Royston & Parmar, 2002). This model can be used with a proportional hazards or a proportional odds framework, and time-varying covariate effects may also be included. Recent papers have used this model with IPTW; although these methods mainly focus on RMS, this is an interesting area for future development (Jackson, 2016; Mozumder et al., 2021).

We found that currently available software for estimation of semiparametric AFT models either did not allow for weights (Pang et al., 2021) or were limited in their available estimands (Chiou et al., 2015, 2014). Therefore, none of the semiparametric AFT models we evaluated met all inclusion criteria for our study, where we focus on methods that are readily accessible to analysts conducting applied studies. However, this is an area of active development, so new methods and software implementations may be available soon.
3.4 Kaplan–Meier method

Another strategy is to avoid parametric assumptions entirely, and use Kaplan–Meier curves with IPTW to estimate the survival functions (Cole & Hernán, 2004; Xie & Liu, 2005):

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{\hat{d}_j}{\hat{n}_j} \right),$$

where $\hat{d}_j = \sum_{i: T_i = t_j} \omega_i \delta_i$ is the weighted number of events and $\hat{n}_j = \sum_{i: T_i \geq t_j} \omega_i$ is the weighted number of people at risk at time $t_j$. Weighted curves are easily estimated in R via the `survival` package.

3.5 Pseudo-observations

Another nonparametric method that can estimate the survival curve is the recently developed pseudo-observations approach (Andersen et al., 2004; Andersen & Pohar Perme, 2010; Andersen et al., 2017). If no missing data (censoring) or confounders are present, the treatment effect can be found by averaging over outcomes for each subject:

$$\psi = E(f(Y)) = \frac{1}{n_Z} \sum_i f(Y_i),$$

where $n_Z$ is the total number of subjects receiving treatment $Z$. When there is censoring, we can define the pseudo-observation for subject $i$ as

$$\hat{\psi}_i = n_Z \hat{\psi} - (n_Z - 1) \hat{\psi}^{-i},$$

where $\hat{\psi}$ is a consistent estimate for $\psi$ and $\hat{\psi}^{-i}$ is the estimator applied to all observations excluding subject $i$. The pseudo-observations are then used in place of all observations (censored and noncensored). In a survival context, $\psi$ is the probability of survival at time $t$, and $\psi_i$ is a binary indicator for the status (alive/dead) of subject $i$ at time $t$. We can estimate $S_Z(t)$ using the Kaplan–Meier method, which then can estimate the pseudo-observations ($\hat{\psi}_i$) for all observations in arm $Z$.

$$\hat{S}_{Z1} = n_Z \hat{S}_Z(t) - (n_Z - 1) \hat{S}_Z(t)^{-i}.$$ 

Heuristically, $n_Z \hat{S}(t)$ is the expected number of patients alive at time $t$, and $(n_Z - 1) \hat{S}(t)^{-i}$ is the expected number of patients other than patient $i$ alive at time $t$. $\hat{\psi}_i$ is used in place of (possibly unobserved) survival status at time $t$. Therefore, the probability of being alive at time $t$ in arm $Z$ is

$$\hat{P}(Y|Z, t) = \frac{1}{n_Z} \sum_i \hat{\psi}_{Z1}.$$

Incorporating propensity-score weights to account for confounding is straightforward.

$$\hat{P}(Y|Z, t) = \frac{1}{n_Z} \sum_i \omega_i \hat{\psi}_{Z1}.$$

There are several benefits of the pseudo-observations approach. It can be used in a causal framework, and requires weaker assumptions than regression-based models. It also readily supports estimation of many outcomes of interest. We can estimate $\hat{P}(Y|Z, t)$ at each failure time, defining the survival curve, which, in turn, allows us to estimate the measures of interest described in Section 3.1. Pseudo-observations can also be used to estimate RMS. This method can be implemented using the R package `pseudo` (Pohar Perme, & Gerster, 2017).
3.6 Variance estimation

To conduct inference for the outcomes of interest, we need to quantify uncertainty. Some methods have closed-form variance estimators for our estimands of interest, but others do not. This motivates the use of a nonparametric bootstrap to estimate variances and find confidence intervals (Efron & Tibshirani, 1993). Bootstrap-based estimates are available for every combination of model and outcome we discuss above. Further, to correctly account for the variance of \( \hat{e} \), the estimated propensity scores, we can reestimate the propensity score within each bootstrap iteration (Austin, 2016). One consideration when using the bootstrap is the number of observations that remain present in the risk set at the longest follow-up times of interest; there must sufficient numbers so that the bootstrap replicates can obtain valid estimates.

4 SIMULATION STUDIES

4.1 Simulation methods

We tested the performance of the methods described above using simulation studies. These studies follow the framework of Morris et al. (2019) for best practices in conducting simulation studies for statistical methods.

**Aims:** This aims of this study were to neutrally assess the performance of IPTW survival methods in observational studies when nonproportional hazards are present.

**Data-generating mechanisms:** We considered seven data-generating mechanisms. We based our covariate distributions, outcome distributions, and effect sizes on the NCDB cancer studies to mimic real-world scenarios as best as possible.

First, we simulated a set of covariates based on the observed multivariate distribution of the NCDB covariates in the renal dataset. We used the R package `genOrd` (Barbiero & Pier, 2015) to draw variables representing gender, age, stage, histology, tumor grade, Charlson comorbidity score, race, ethnicity, insurance status, facility type, income, and education.

As our goal here is to evaluate the survival analysis methods, and not the propensity score estimation methods, we held this set of covariates constant for each simulation; we employ simulated covariates instead of the actual covariates so that our data and simulation code can be shared. (Available online at https://github.com/BethHandorf/NonPH_IPTW)

Next, we used a propensity score model to determine each subject’s probability of being assigned to the treatment condition (versus control). The effect size (log-scale) for each covariate ranged from 0.8 (stage) to 0.03 (1 year of age). (See Online Appendix Table A1 for full specification.) Effect sizes were chosen to emulate realistic scenarios grounded in the NCDB studies. We set the intercept such that, on average, half of the population received treatment. In each simulation, observed treatment status was drawn from the binomial based on each individuals’ treatment probability, as defined by our propensity score model.

In our base case, survival times were drawn based on a Weibull model with a time-varying treatment effect.

\[
h(t) = \gamma \lambda t^{\gamma-1} \exp(\beta_1 Z + f(Z, t) + X\psi). \tag{2}\]

In our base case, the HR varied by log-time \( f(Z, t) = \beta_2 Z \log(t) \). We set \( \beta_1 = -0.69, \beta_2 = 0.25 \), with a sample size of 5000 total cases. For each individual, survival times under \( Z = 1 \) and \( Z = 0 \) were both drawn using the R package `simsurv` (Brilleman, 2019). In additional scenarios, we generated survival times using the generalized gamma AFT model, with times drawn using the R package `flexsurv`

\[
\log(T_i) = \mu + \beta_1 Z_i + X\psi + \sigma \epsilon_i, \tag{3}\]

\[
\epsilon_i \sim \left( \frac{|\lambda|}{\epsilon_i \Gamma(\lambda+2)} \right) \lambda^{-1} \epsilon_i^{\lambda-1} \exp(-\lambda^{-2} \epsilon_i^2),
\]

with \( \beta_1 = -0.69 \) and \( \sigma = 1.2 \). We varied the values of \( \lambda \) and \( \mu \) in different scenarios.

In each survival model, covariate effect sizes ranged from 0 to 0.8 (see Online Appendix Table A1). Censoring times were drawn from the uniform (with a range of 8–15), to mimic administrative censoring. Observations with a censoring time less than their true event time were censored at that time.
We simulated data using seven parametric survival models, varying equations (2) and (3):

1. Base case: Weibull model with $\beta_1 = -0.69$; $f(Z, t) = 0.25Z\log(t)$
2. Weibull model with PWC HR: $\beta_1 = 0$; $f(Z, t) = -0.25ZI(t \geq 2)$, where $I(\cdot)$ is the indicator function
3. Base case with modest nonproportionality $\beta_1 = -0.69$; $f(Z, t) = 0.125Z\log(t)$
4. Base case with modest treatment effect $\beta_1 = -0.41$; $f(Z, t) = 0.25Z\log(t)$
5. Base case, with a smaller sample size ($N = 500$ per arm)
6. Generalized gamma survival model with $\mu = 2$ and $\lambda = -0.5$
7. Generalized gamma model with $\mu = 3.5$ and $\lambda = 2.5$

Scenario 1 (base case) had a large treatment effect with substantial nonproportionality. Scenarios 2–4 approximated the parameters and nonproportionality from the clinical examples, and scenario 5 was used to assess how the methods perform with a more modest sample size. The generalized gamma models used in scenarios 6–7 both result in survival curves with nonproportional hazards, but they have different forms of nonproportionality from scenarios 1–5, and from each other.

**Estimands:** Outcomes of interest for this study were median survival, RMS, and survival probability at time $T$, as described in Section 3.1.

**Methods:** The seven data-generating mechanisms were each used to generate 500 simulated datasets. For each patient, we generated three simulated survival times: (1) their survival time when $Z = 1$, (2) their survival time if $Z = 0$, and (3) their observed survival time, subject to their treatment assignment and their censoring time. The first two outcomes were used to calculate the true treatment effects for each patient, which, in turn, were aggregated to calculate the treatment effects for the population. The final outcome (along with a censoring indicator) was used to fit the survival models.

For each set of simulated outcomes, we first estimated the propensity scores using a logistic regression model, and then calculated the IPTWs. Using the estimated weights, we fit the models described in Sections 3.2–3.5. Variance estimates, along with 95% confidence intervals, were calculated via a nonparametric bootstrap, as described in Section 3.6, based on 500 bootstrap replicates.

**Performance measures:** We assessed the performance of each model based on the bias, coverage, and bootstrap standard error (SE). Bias was defined as the difference between the estimated treatment effect and the true treatment effect, and we calculated the mean bias over the 500 simulated samples. For clarity, we present the absolute value of the mean bias in the results below. A bias close to zero is desirable for each method/estimand combination. Coverage was defined as the proportion of simulations where the estimates fall within their respective 95% confidence intervals. Coverage close to the nominal 0.95 level is desirable. SE is the average bootstrap-based SE over the 500 simulations. Smaller SEs are desirable, as they indicate more efficient estimates.

### 4.2 Simulation results

For the base case, the bias was often largest for the standard Cox model, as expected given the substantial degree of nonproportionality in the simulation. For example, considering median survival, bias for the Cox model was $-0.347$ years, whereas all other methods had biases less than $-0.069$. The generalized gamma AFT model also had large bias relative to other methods; at 10 years, its bias was actually greater than that of the Cox model (2.5% vs. 2.0%, respectively). Even though this three-parameter model is flexible and allows for nonproportionality, fitting this misspecified parametric model still introduced substantial biases into the results. In the base case, the Cox model with a treatment effect varying by log-time was correctly specified, as was the fully parametric AFT model with variable location and scale parameters. These models had lower biases than the standard Cox model (e.g., for median survival, biases were $-0.042$ and $-0.056$, respectively). However, the nonparametric Kaplan–Meier and pseudo-observation methods generally had the lowest biases (e.g., biases for median survival were $-0.021$ and $-0.020$, respectively). Finally, we note that the RMS was less sensitive to the choice of method than the other outcomes of interest; the relative differences in the sizes of the biases were smallest for this survival outcome (see Figure 1).

As shown in Table 1, coverage was worst for the standard Cox model, followed by the generalized gamma AFT model. The time-varying Cox models did better, although coverage was surprisingly low for the log-time model at the 2-year outcome (0.82). Coverage for the nonparametric methods was close to the nominal level for all outcomes. SEs of the differences in treatment effects were usually slightly larger for the nonparametric methods than for the parametric and
**FIGURE 1** Mean bias\(^a\) in estimates from simulation studies

CTV, Cox time-varying; LT, log-time; PWC, piece-wise constant; AFT, accelerated failure time; GG, generalized gamma; WBLLS, Weibull location scale; Pseudo, pseudo-observations; Wtd KM, weighted Kaplan–Meier; Mod NPH, modest nonproportional hazards; Mod TE, modest treatment effect; Sm SS, small sample size.

\(^a\)For ease of comparison, we present the absolute value of the mean bias.

**TABLE 1** Empirical coverage probability of the 95% bootstrap percentile confidence interval

| Method   | 2 Years | 5 Years | 10 Years | Median | RMS   |
|----------|---------|---------|----------|--------|-------|
| Cox      | 0.00    | 0.07    | 0.61     | 0.81   | 0.71  |
| CTV LT   | 0.82    | 0.91    | 0.93     | 0.94   | 0.93  |
| CTV PWC  | 0.96    | 0.96    | 0.95     | 0.95   | 0.93  |
| AFT GG   | 0.00    | 0.37    | 0.44     | 0.94   | 0.90  |
| AFT WBL LS | 0.89   | 0.93    | 0.93     | 0.95   | 0.93  |
| Pseudo   | 0.97    | 0.97    | 0.96     | 0.95   | 0.94  |
| Wtd KM   | 0.97    | 0.96    | 0.96     | 0.95   | 0.94  |

semiparametric methods (see Online Appendix Figure A1). Comparing the SEs for the Kaplan–Meier versus the Weibull model, the increases ranged from 0.3–24.4%, and were lowest for RMS and highest for 2-year survival.

For PWC hazards (scenario 2), biases were typically lowest for the nonparametric methods and when using a PWC time-varying Cox model. An exception to this was for RMS, where the naïve Cox model and Weibull AFT model had the lowest bias. This likely occurred because bias from different timepoints was in different directions (positive at 2 and 5 years, negative at 10 years), and may have canceled each other out when calculating the area under the survival curve. (Note that Figure 1 shows the absolute value of the average biases). Coverages (see Online Appendix Table A2) were close to the nominal when biases were low. SEs were generally similar to be base case, except for median survival, where they were somewhat larger.

When the effect of nonproportionality was modest (scenario 3), biases tended to be smaller than those of the base case, especially for estimates of survival probabilities at a given timepoint. When the treatment effect was modest (scenario 4),
results were generally similar to those of the base case. In the small sample size case (scenario 5), the bias was similar to that of the base case, with the exception of RMS, where across methods, biases were larger than those of the base case. As expected, SEs were higher when the sample size was smaller.

Scenarios 6 and 7, which used the generalized gamma distribution, had low biases when using generalized gamma parametric models. This is expected, as the models are correctly specified (unlike in scenarios 1–5). Biases were also low when using nonparametric models, or the Cox model with a PWC time-varying hazard. In scenario 6, when \( \lambda = -0.5 \), the standard Cox model and the Weibull AFT model performed particularly poorly, while when \( \lambda = 2.5 \), the Cox model with a treatment effect varying by log-time generally performed the worst. Models with large biases also had the worst coverage. Notably, many methods applied to scenarios 6 and 7 resulted in high coverage, and coverage was above the nominal in some cases (especially for the nonparametric methods), indicating that the SEs may be overestimated by the bootstrap procedure. Similar to what we saw in the base case, SEs were lowest when the correctly specified parametric models were used, but improvements over the nonparametric methods were small.

5 | COMPARATIVE EFFECTIVENESS OF CANCER THERAPIES

We applied each method to estimate the effect of treatments on survival in two clinical studies using the National Cancer Database (NCDB). The NCDB is a large registry that encompasses approximately 70% of newly diagnosed cancer cases in the United States, containing patient, tumor, and treatment measures (Boffa et al., 2017). The studies discussed here featured two different forms of nonproportional hazards. In the first, the treatment groups had similar survival initially, but differences arose in longer term outcomes. In the second example, the benefit of the treatment was attenuated over time. Both studies used deidentified NCDB data, which has been determined by our Institutional Review Board to be not human subjects research.

5.1 | Sarcoma

Soft-tissue sarcoma is a rare cancer of the connective tissue (e.g., fat, muscle, blood vessel), usually arising in the extremities. Stage III disease (high grade, large, or deep tumor) is typically treated surgically or with radiation therapy. It remains uncertain whether chemotherapy provides an overall survival benefit for these patients, and its use is optional according to national guidelines (von Mehren, 2020). This question was studied in an NCDB cohort of Stage III sarcoma patients treated with definitive surgical resection of the primary tumor, with or without chemotherapy pre-/postoperatively. Of the 5337 cases, 28% were treated with chemotherapy (Movva et al., 2015).

Due to the nonrandomized nature of the data, there were substantial differences between treatment groups. For example, patients treated with chemotherapy were younger and had fewer recorded comorbid conditions. Furthermore, the hazards for the treatment groups exhibited substantial nonproportionality, as is clearly evident upon examining the complementary log-log survival curves (see Figure 2). Testing the Schoenfeld residuals provides further evidence of nonproportionality \((p = 0.047)\). We used IPTW to account for covariate imbalance between the treatment groups. The propensity score was estimated using all available covariates (including age, gender, insurance status, income, comorbidity score, tumor histology, grade, size, anatomic site, treating facility type, and travel distance.) The IPTW Kaplan–Meier curves are shown in Figure 2. There was little effect of chemotherapy during the first 2 years, but patients treated with chemotherapy did better in the long term. These results motivated simulation scenario 2 (PWC). We estimated propensity scores via logistic regression. Upon assessing balance, the maximum standardized difference dropped from 0.64 (age) to 0.11 (histology). All other variables had standardized differences <0.1, with most being <0.05, indicating good balance.

We applied each method to the sarcoma data. Results for RMS, 2-year survival, and 5-year survival are shown in Table 2 (sample sizes were not sufficient at the end of follow-up to obtain bootstrap confidence intervals for median and 10-year survival). The different methods yielded different effect size estimates; however, these differences were often modest and in some cases would not change statistical inferences. For example, \( \Delta \text{RMS} \) varied from a minimum of 0.409 (AFT with a generalized gamma distribution) to a maximum of 0.468, but all models showed an improvement in RMS for patients treated with chemotherapy. In other cases, the inferences would differ based on the model chosen. At 2 years, the standard Cox model showed a statistically significant difference of 3.6% between arms. Both AFT models also found significantly higher 2-year survival in the chemotherapy group. However, the weighted Kaplan–Meier method found a nonsignificant difference of –0.7%. Both time-varying Cox models and the pseudo-observations method also produced nonsignificant differences.
We note that in this data analysis, we found that the pseudo-observations method was sensitive to the mean of the estimated weights. While in expectation, the mean of the weights is 1, in practice, there may be small deviations. The pseudo-observations method was more affected by these deviations than the other methods we evaluated. This sensitivity can be explained by how the weights are applied and how the survival function is estimated. Each patient’s (binary) contribution to the survival function is weighted, and the weighted indicators are then averaged. Therefore, if the mean of the weights is not equal to 1, at baseline, the survival function will not start at 1. To address this issue, we centered the weights, so their mean was exactly 1. However, we note that this was done heuristically, and not theoretically driven. Further exploration of this issue is warranted.

### 5.2 Renal cancer

The second motivating clinical example for this work was a study of surgical options for early-stage renal cancer. These small tumors can be treated with radical nephrectomy (RN) or partial nephrectomy (PN). PN preserves more kidney
tissue, but surgery/recovery is more difficult and it is more likely to be given to healthier patients (Motzer, 2020; Ristau et al., 2018). Here, we focus on a subgroup: patients aged 51–60 with T1a tumors, which gives a sample size of 28,973 patients (61.1% treated with PN). In unadjusted analysis, patients given PN had improved survival over those given RN. However, these results are subject to confounding by indication. Examination of the complementary log-log plots shows a notable deviation from nonproportionality, which is confirmed by testing the Schoenfeld residuals ($p<0.0001$) (Figure 3) (Grambsch & Therneau, 1994). Again, we estimated propensity scores via logistic regression. Here, the maximum standardized difference dropped from 0.30 (facility type) to 0.01 (geographic location), indicating excellent balance.

When we fit an IPTW Cox model to these data, allowing the hazards to vary as a function of log-time, such that $h(t) = a_0(t) \exp(\beta_1 Z + \beta_2 \log(t))$ the maximum likelihood estimates are $\hat{\beta}_1 = -0.468$ and $\hat{\beta}_2 = 0.136$. Therefore, the non-proportionality observed here is close to our modest time-varying hazard simulation scenario 3, and the treatment effect is modest as well, similar to what we used in scenario 4.

Our results were largely similar, regardless of what model was used (see Table 3). This was particularly the case for the RMS and 5-year survival. Two-year survival showed the largest range in results, with estimated differences ranging from 0.9% (generalized gamma AFT model) to 1.6% (Cox model with piece-wise time-varying effects). The standard Cox model

---

**Figure 3** Survival in early-stage renal cancer by surgery type

**Table 3** Effect of PN versus RN in renal cancer: Results from each model

|                | RMS  | LCL  | UCL  | 2 Years | LCL  | UCL  | 5 Years | LCL  | UCL  |
|----------------|------|------|------|---------|------|------|---------|------|------|
| Cox            | Δ    | LCL  | UCL  | Δ       | LCL  | UCL  | Δ       | LCL  | UCL  |
| TV Cox: log-T  | 0.267| 0.199| 0.330| 0.014   | 0.010| 0.018| 0.029   | 0.022| 0.036|
| TV Cox: PWC    | 0.264| 0.196| 0.327| 0.016   | 0.012| 0.020| 0.030   | 0.021| 0.037|
| AFT Gen Gamma  | 0.263| 0.194| 0.326| 0.009   | 0.007| 0.012| 0.026   | 0.020| 0.033|
| AFT Wbl TV shape | 0.264| 0.196| 0.329| 0.019   | 0.014| 0.023| 0.031   | 0.024| 0.038|
| Weighted K-M   | 0.268| 0.200| 0.333| 0.015   | 0.011| 0.020| 0.029   | 0.020| 0.036|
| Pseudo-obsa    | 0.269| 0.200| 0.334| 0.015   | 0.011| 0.020| 0.029   | 0.020| 0.036|

aCentered weights.
Δ = Difference between Treatment Effects for PN vs RN, boldface denotes significance.
LCL = Lower confidence limit, UCL = Upper confidence limit.
gave an estimated difference of 1%, compared to 1.5% from the Kaplan–Meier method. Even with these small effects, we found that none of the 95% confidence limits crossed zero, so all models would lead to the same inference: a small benefit for PN over RN. The congruous inferences were partially driven by the very large sample size, and the resulting small confidence intervals; however, given the clear evidence of nonproportionality, it is notable that the point estimates for the standard Cox model were close to those of time-varying or nonparametric methods.

6  | DISCUSSION

When analyzing survival outcomes, we often see nonproportional hazards. Based on the findings of our neutral comparison study, it seems that nonproportionality is readily addressable in practice with weighted Kaplan–Meier curves, the simplest method we assessed. It performed quite well across a variety of scenarios and outcome measures of interest. The IPTW Kaplan–Meier method does not require specialized software or methods. There is an efficiency penalty, but the increase in the size of the SEs (compared to parametric methods) was small in practice. We believe that the larger SEs are a minor drawback when compared to the benefits of fewer assumptions and reduced bias.

This work has several limitations. Here, we used simple IPTW weights. Alternative propensity score weights have also been proposed (Austin & Stuart, 2015) notably, variance stabilized weights have good properties, especially when one treatment is given to a small proportion of the patients. In the current paper, we use a correctly specified logistic regression model estimate the propensity score. The impact of model misspecification was out the scope of this paper, and this issue has been studied elsewhere (Austin & Stuart, 2017). In modern practice, propensity score models can be estimated using flexible approaches that do not rely on a specific parametric form (van der Laan et al., 2007; Pirracchio et al., 2015; Cannas & Arpino, 2019; Zou et al., 2021). Such methods have been shown to improve prediction, and may help to achieve balance between treatment arms. A further limitation of our paper is that we did not consider potential effects of unmeasured confounders; propensity score approaches will not balance unmeasured confounders (Austin, 2011). Additionally, the nonproportionality we studied was limited to four functional forms. Our simulation studies were directly informed by the two cancer studies discussed in Section 5. We extended our simulations by examining survival data generated by two very different generalized gamma models, and found consistent results. However, it is possible that some forms of nonproportionality may behave differently.

In this paper, we sought to better understand how choice of method may affect results of these real-world analyses, and to learn broader lessons about how best to accommodate nonproportionality in large, observational studies. Although both clinical examples had large sample sizes, we found in simulations that the results held up even when sample sizes were small. Taken together, our simulation results and clinical examples show that one can easily protect against incorrect inferences using IPTW Kaplan–Meier curves to estimate treatment effects.

ACKNOWLEDGMENT
Elizabeth Handorf’s research in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA006927 and U54CA221705. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The data used in the study are derived from a deidentified NCDB file. The American College of Surgeons and the Commission on Cancer have not verified and are not responsible for the analytic or statistical methodology employed, or the conclusions drawn from these data by the investigator.

CONFLICT OF INTEREST
The authors declare that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT
All codes to generate and analyze data for the simulation studies are available at https://github.com/BethHandorf/NonPH_IPTW
All codes to analyze the data for the clinical examples are also available on github. Applications to obtain the data used here can be made to the American College of Surgeons at https://www.facs.org/quality-programs/cancer/ncdb/puf

IRB Statement: As this study used only deidentified data, it was determined by the Fox Chase Cancer Center Institutional Review Board to be not human subjects research.
This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge “Reproducible Research” for making publicly available the code necessary to reproduce the reported results. The results reported in this article were reproduced partially due to data privacy reasons.

ORCID
Elizabeth A. Handorf https://orcid.org/0000-0003-0445-8978

REFERENCES
Ambrogi, F., Iacobelli, S., & Andersen, P. K. (2022). Analyzing differences between restricted mean survival time curves using pseudo-values. BMC Medical Research Methodology, 22(1), 71.
Andersen, P. K., Hansen, M. G., & Klein, J. P. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. Lifetime Data Analysis, 10(4), 335–350.
Andersen, P. K., & Pohar Perme, M. (2010). Pseudo-observations in survival analysis. Statistical Methods in Medical Research, 19(1), 71–99.
Andersen, P. K., Syriopoulou, E., & Parner, E. T. (2017). Causal inference in survival analysis using pseudo-observations. Statistics in Medicine, 36(17), 2669–2681.
Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399–424.
Austin, P. C. (2016). Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. Statistics in Medicine, 35(30), 5642–5655.
Austin, P. C., & Stuart, E. A. (2015). 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(28), 3661–3679.
Austin, P. C., & Stuart, E. A. (2017). 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(4), 1654–1670.
Barbiero, A., & Pier, A. F. (2015). GenOrd: Simulation of discrete random variables with given correlation matrix and marginal distributions.
Bofa, D. J., Rosen, J. E., Mallin, K., Loomis, A., Gay, G., Palis, B., Thoburn, K., Gress, D., McKellar, D. P., Shulman, L. N., Facktor, M. A., & Winchester, D. P. (2017). Using the national cancer database for outcomes research: A review. JAMA Oncology, 3(12), 1722–1728.
Brilleman, S. (2019). Simsurv: Simulate Survival Data.
Buckley, J., & James, I. (1979). Linear regression with censored data. Biometrika, 66(3), 429–436.
Burzykowski, T. (2022). Semi-parametric accelerated failure-time model: A useful alternative to the proportional-hazards model in cancer clinical trials. Pharmaceutical Statistics, 21(2), 292–308.
Cannas, M., & Arpino, B. (2019). A comparison of machine learning algorithms and covariate balance measures for propensity score matching and weighting. Biometrical Journal, 61(4), 1049–1072.
Chiou, S., Kang, S., & Yan, J. (2015). Rank-based estimating equations with general weight for accelerated failure time models: An induced smoothing approach. Statistics in Medicine, 34(9), 1495–1510.
Chiou, S. H., Kang, S., & Yan, J. (2014). Fitting accelerated failure time models in routine survival analysis with R Package aftgee. Journal of Statistical Software, 61, 1–23.
Cole, S. R., & Hernán, M. A. (2004). Adjusted survival curves with inverse probability weights. Computer Methods and Programs in Biomedicine, 75(1), 45–49.
Collett, D. (2015). Modelling survival data in medical research. Chapman and Hall/CRC.
Conner, S. C., Sullivan, L. M., Benjamin, E. J., LaValley, M. P., Galea, S., & Trinquart, L. (2019). Adjusted restricted mean survival times in observational studies. Statistics in Medicine, 38(20), 3832–3860.
Cox, C., Chu, H., Schneider, M. F., & Muñoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. Statistics in Medicine, 26(3), 4353–4374.
Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202.
Deb, S., Austin, P. C., Tu, J. V., Ko, D. T., Mazer, C. D., Kiss, A., & Freams, S. E. (2016). A review of propensity-score methods and their use in cardiovascular research. Canadian Journal of Cardiology, 32(2), 259–265.
Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman and Hall/CRC.
Enewold, L., Parsons, H., Zhao, L., Bott, D., Rivera, D. R., Barrett, M. J., Virnig, B. A., & Warren, J. L. (2020). Updated overview of the seer-medicare data: Enhanced content and applications. Journal of the National Cancer Institute. Monographs, 2020(55), 3–13.
Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. American Journal of Epidemiology, 173(7), 761–767.
Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3), 515–526.
Hernán, M. A., Cole, S. R., Margolick, J., Cohen, M., & Robins, J. M. (2005). Structural accelerated failure time models for survival analysis in studies with time-varying treatments. Pharmacoepidemiology and Drug Safety, 14(7), 477–491.
Horiguchi, M., Hassett, M. J., & Uno, H. (2019). How do the accrual pattern and follow-up duration affect the hazard ratio estimate when the proportional hazards assumption is violated? *The Oncologist*, 24(7), 867–871.

Jackson, C. (2016). Flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software*, 70(8), 1–33.

Jiménez, J. L. (2022). Quantifying treatment differences in confirmatory trials under non-proportional hazards. *Journal of Applied Statistics*, 49(2), 466–484.

Jin, Z., Lin, D. Y., & Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, 93(1), 147–161.

Kalbfleisch, J. D., & Prentice, R. L. (1981). Estimation of the average hazard ratio. *Biometrika*, 68(1), 105–112.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457–481.

Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*, 2nd ed. Springer.

Komárek, A., Lesaffre, E., & Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, 14(3), 726–745.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine*, 23(19), 2937–2960.

Mao, H., Li, L., Yang, W., & Shen, Y. (2018). On the propensity score weighting analysis with survival outcome: Estimands, estimation, and inference. *Statistics in Medicine*, 37(26), 3745–3763.

Marshall, A. W., & Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, 84(3), 641–652.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102.

Motzer, R. J. (2020). NCCN clinical practice guidelines in oncology (NCCN Guidelines®) Kidney Cancer version 1.2021.

Movva, S., von Mehren, M., Ross, E. A., & Handorf, E. (2015). Patterns of chemotherapy administration in high-risk soft tissue sarcoma and impact on overall survival. *Journal of the National Comprehensive Cancer Network*, 13(11), 1366–1374.

Mozumder, S. I., Rutherford, M. J., & Lambert, P. C. (2021). Estimating restricted mean survival time and expected life-years lost in the presence of competing risks within flexible parametric survival models. *BMC Medical Research Methodology*, 21(1), 52.

Pang, M., Platt, R. W., Schuster, T., & Abrahamowicz, M. (2021). Flexible extension of the accelerated failure time model to account for nonlinear and time-dependent effects of covariates on the hazard. *Statistical Methods in Medical Research*, 30(11), 2526–2542.

Pirracchio, R., Petersen, M. L., & van der Laan, M. (2015). Improving propensity score estimators’ robustness to model misspecification using super learner. *American Journal of Epidemiology*, 181(2), 108–119.

Pohar Perme, M., & Gerster, M. (2017). Pseudo: Computes pseudo-observations for modeling.

Rahman, R., Fell, G., Ventz, S., Arfé, A., Vanderbeek, A. M., Trippa, L., & Alexander, B. M. (2019). Deviation from the proportional hazards assumption in randomized phase 3 clinical trials in oncology: Prevalence, associated factors, and implications. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, 25(21), 6339–6345.

Ristau, B. T., Handorf, E. A., Cahn, D. B., Kutikov, A., Uzzo, R. G., & Smaldone, M. C. (2018). Partial nephrectomy is not associated with an overall survival advantage over radical nephrectomy in elderly patients with stage Ia-II renal masses: An analysis of the National Cancer Data base. *Cancer*, 124(19), 3839–3848.

Robins, J., & Tsiatis, A. A. (1992). Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*, 79(2), 311–319.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 412–8.

Rosenbaum, P. R., & Rubin, D. B. (1984). Bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516–524.

Royston, P., & Parmar, M. K. B. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, 21(15), 2175–2197.

Royston, P., & Parmar, M. K. B. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine*, 30(19), 2409–2421.

Sachs, M. C., & Gabriel, E. E. (2022). Event history regression with pseudo-observations: Computational approaches and an implementation in R. *Journal of Statistical Software*, 102, 1–34.

Scheepers, M., Wakounig, S., & Heinze, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Statistics in Medicine*, 28(19), 2473–2489.

Therneau, T. M. (2020). A Package for Survival Analysis in R.

Tian, L., Jin, H., Uno, H., Lu, Y., Huang, B., Anderson, K. M., & Wei, L. (2020). On the empirical choice of the time window for restricted mean survival time. *Biometrics*, 76(4), 1157–1166.

Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M., & Wei, L.-J. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32(22), 2380–2385.

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1).

von Mehren, M. (2020). NCCN Clinical practice guidelines in Oncology (NCCN Guidelines®) Soft Tissue Sarcoma Version 2.2020.

Wei, L. J. (1992). The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15), 1871–1879.
Xie, J., & Liu, C. (2005). Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine, 24*(20), 3089–3110.

Zhang, M., & Schaubel, D. E. (2012). Double-robust semiparametric estimator for differences in restricted mean lifetimes in observational studies. *Biometrics, 68*(4), 999–1009.

Zou, B., Mi, X., Tighe, P. J., Koch, G. G., & Zou, F. (2021). On kernel machine learning for propensity score estimation under complex confounding structures. *Pharmaceutical Statistics, 20*(4), 752–764.

**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:** Handorf, E. A., Smaldone, M. C., Movva, S., & Mitra, N. (2024). Analysis of survival data with nonproportional hazards: A comparison of propensity-score-weighted methods. *Biometrical Journal, 66*, 202200099. [https://doi.org/10.1002/bimj.202200099](https://doi.org/10.1002/bimj.202200099)