A sensitivity analysis approach for the causal hazard ratio in randomized and observational studies

Rachel Axelrod | Daniel Nevo

Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv, Israel

Correspondence
Rachel Axelrod, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 6997801, Israel. Email: axelrod1@mail.tau.ac.il

Funding information
Israel Science Foundation, Grant/Award Number: ISF grant No. 827/21

Abstract
The hazard ratio (HR) is often reported as the main causal effect when studying survival data. Despite its popularity, the HR suffers from an unclear causal interpretation. As already pointed out in the literature, there is a built-in selection bias in the HR, because similarly to the truncation by death problem, the HR conditions on post-treatment survival. A recently proposed alternative, inspired by the Survivor Average Causal Effect, is the causal HR, defined as the ratio between hazards across treatment groups among the study participants that would have survived regardless of their treatment assignment. We discuss the challenge in identifying the causal HR and present a sensitivity analysis identification approach in randomized controlled trials utilizing a working frailty model. We further extend our framework to adjust for potential confounders using inverse probability of treatment weighting. We present a Cox-based and a flexible non-parametric kernel-based estimation under right censoring. We study the finite-sample properties of the proposed estimation methods through simulations. We illustrate the utility of our framework using two real-data examples.

KEYWORDS
causal inference, frailty models, survival analysis

1 | INTRODUCTION
The analysis of time-to-event data, also known as survival analysis, is one of the backbones of clinical and epidemiological research (Andersen et al., 2012, 2021). When studying the causal effect of a treatment or exposure on a time-to-event outcome, arguably the most ubiquitously reported measure is the hazard ratio (HR). At each time \( t \), the HR compares the outcome rate between the treated and untreated, among the event-free individuals. A commonly-employed tool for HR estimation is the Cox proportional hazard (PH) model (Cox, 1972), which assumes that the HR is constant through time. To make a fair comparison, researchers seek to conduct randomized controlled trials (RCTs) within which treatment is assigned randomly. Alternatively, in observational studies, when the treatment or exposure assignment is not randomized, researchers collect rich data and spend considerable time selecting the appropriate covariates and models for their analyses.

In the past decade, it has become increasingly clear that a causal interpretation of the HR is unlikely (Aalen et al., 2015; Hernán, 2010; Martinussen et al., 2020). As first discussed by Hernán (2010), there is a built-in selection bias in the HR as a parameter which prevents a causal interpretation. While the treated and untreated are comparable at baseline, either by design or conditionally on covariates,
among those who survived until time $t$, if the treatment is beneficial, then the treated are expected to be on average more frail than the untreated as a result of lower mortality among the treated up to time $t$.

Although alternatives to the HR were suggested, it remains the most popularly reported measure even in studies targeting causal effects implicitly or explicitly (Assel et al., 2019; Lang & Altman, 2015). Therefore, having an alternative HR-like measure that admits a causal interpretation will be useful in practice. As part of teasing out the complexity of the interpretation of HRs, Martinussen et al. (2020) introduced a causal HR defined within the subpopulation of those who will potentially survive until time $t$ regardless of the treatment received. This definition of the causal effect is similar to the survivor average causal effect (SACE) (Rubin, 2006). The SACE is the effect of treatment on a non-survival outcome among the sub-population that would have survived under either treatment arm. The SACE is not point-identified under standard assumptions, prompting researchers to often center their efforts around sensitivity analysis methods (Hayden et al., 2005; Zehavi & Nevo, 2021).

While the causal HR was previously defined (Martinussen et al., 2020), the framework by which a researcher should follow in order to identify and estimate the causal HR from real data has not been fully established. In this paper, we present a detailed identification strategy for the causal HR in randomized studies. We develop a sensitivity analysis and couple it with two possible estimators. We extend our approach to account for observed confounders and propose inverse probability of treatment weighting (IPTW) estimators. Our sensitivity analysis is built upon the idea of frailty models, accounting for the imbalance between the treatment groups due to unobserved factors that affect survival and cannot be adjusted for.

The rest of the paper is organized as follows. In Section 2, we review two motivating examples, an RCT and an observational study. In Section 3, we present the notations and assumptions, and we clarify the motivation for targeting the causal HR. In Section 4, we provide identification formulas and propose a sensitivity analysis approach stemming from these formulas. In Section 5, we present a kernel-based and a Cox-based estimation methods in randomized studies and extend the approach to observational studies using IPTW. Evaluation of this estimators using a simulation study is presented in Section 6. Finally, in Section 7, we return to the two motivating examples and apply our proposed methodology.

## 2 TWO MOTIVATING EXAMPLES

Our first motivating example is the Urothelial Carcinoma data (Powles et al., 2018). These data are from the IMvigor211 RCT study, which aimed to evaluate the effectiveness of an immunotherapy treatment, Atezolizumab (Atezo), in comparison to the standard care of chemotherapy treatment (Chemo) in patients with locally advanced or metastatic urothelial carcinoma. The study outcome was the survival time defined as the time between patient’s study enrollment and death. The data available for us (Gorfine et al., 2020) included the 625 patients from the IC1/2/3 subgroup (Powles et al., 2018), out of which 316 (51%) patients were treated with Atezo and 309 (49%) with Chemo.

From the Kaplan–Meier curves in Figure 1, it seems that while at early times the Atezo patients are at higher risk than the Chemo patients, at later times the opposite is true. Furthermore, consider a Cox model analysis that partitioned the time axis into three segments: 0–4 months; 4–8 months; and 8–12 months, and estimated a constant HR within each segment. The estimated HR for the 0–4 and 4–8 months periods is larger than one, although insignificant ($\hat{HR} = 1.34, 95\%$ confidence interval (CI): 0.96–1.88, and $\hat{HR} = 1.09, 95\%$ CI: 0.76–1.57, respectively). However, the estimated HR for the 8–12 months period is significantly lower than one ($\hat{HR} = 0.71, 95\%$ CI: 0.53–0.95), indicating that the Atezo treatment may have a long-term preventive effect. Nevertheless, reporting three HRs instead of a single HR does not alleviate the problem of using HRs as a measure of the causal effect.

Our second motivating example is the Données Informatisées et VAlidées en Transplantation (DIVAT), an ongoing prospective cohort study following kidney transplant recipients. Our goal is to study whether donor’s characteristics (expanded vs. standard donor criteria) influence the recipient’s survival time. As this is an observational study, the two donor type groups are not balanced even at the beginning of the study.

![Survival curves estimated by Kaplan–Meier estimator for urothelial carcinoma patients according to the treatment arm: Atezolizumab (Atezo) vs. chemotherapy (Chemo). Cases later than 12 months were censored at 12 months due to low number of events after this time.](image-url)
3. NOTATIONS, ASSUMPTIONS AND CAUSAL ESTIMANDS

3.1. Notations and assumptions

We employ the potential outcomes framework (Imbens & Rubin, 2015). Let $A$ denote the binary treatment and $T^{A=a,C=\infty}$ denote the potential event time had a patient was assigned to treatment group $a$, $a = 0, 1$, and there was no loss to follow up or administrative censoring, namely that the censoring time was $C = \infty$. This notion of the outcome had there was no censoring aligns with the survival analysis literature that often defines the event time had there was no censoring, although typically without using the potential outcomes framework explicitly. Let $C^{A=a}$ denote the potential censoring time under treatment level $A = a$. Throughout the paper, we assume the standard independent censoring assumption.

**Assumption 1.** Independent censoring. $C^{A=a} \perp T^{A=a,C=\infty}$, for $a = 0, 1$.

Let $X^{A=a} = \min\{T^{A=a,C=\infty}, C^{A=a}\}$ and $\delta^{A=a} = I[T^{A=a,C=\infty} \leq C^{A=a}]$ represent the observed time and the event indicator, respectively, had we intervened and set $A = a$. We also denote the potential outcomes’ event counting processes by $N^{A=a}(t) = \delta^{A=a}I\{X^{A=a} \leq t\}$, and let

$$\lambda^{A=a,C=\infty}(t) := \lambda^{A=a}(t) = \lim_{dt \to 0} (dt)^{-1} \Pr \left( t \leq T^{A=a,C=\infty} < t + dt \mid T^{A=a,C=\infty} \geq t \right) \tag{1}$$

be the hazard rate in terms of the potential outcomes, and let $\lambda^{A=a}(t|Q)$ be the analogous quantity given an event $Q$.

In addition to Assumption 1, we also make the following standard assumptions.

**Assumption 2.** Stable Unit Treatment Value Assumption (SUTVA). There is no interference between patients, and there are no multiple versions of each treatment value $a$ leading to a different outcome.

**Assumption 3.** Randomization. $A \perp X^{A=a}$, for $a = 0, 1$.

For observational studies, we replace Assumption 3 with the following weaker assumption.

**Assumption 4.** Conditional exchangeability. For a vector of observed confounders $Z$, $A \perp X^{A=a} \mid Z$, for $a = 0, 1$.

From Assumption 2, it follows that $X = (1 - A)X^{A=0} + AX^{A=1}$, $\delta = (1 - A)\delta^{A=0} + A\delta^{A=1}$ and $N(t) = (1 - A)N^{A=0}(t) + AN^{A=1}(t)$. Under the randomized trial setting, the observed data for each patient $i$, $i = 1, \ldots, n$, is $(A_i, X_i, \delta_i)$. In the presence of confounders, we assume that we additionally observe $Z_i$. From $X_i$ and $\delta_i$ one can obtain the risk indicator $Y_i(t) = I(X_i \geq t)$, the counting process $N_i(t)$, and the indicator of the jump in the process $N_i(t)$ during the interval $[t, t + dt)$, denoted by $dN_i(t)$. Finally, let $N_i(t|A = a) = N_i(t|A_i = a)$, $dN_i(t|A = a) = dN_i(t|A_i = a)$, and $Y_i(t|A = a) = Y_i(t|A_i = a)$ be the analogous quantities in each treatment arm.

3.2. The causal and the non-causal hazard ratio

Let $HR^{PO}(t) = \frac{\lambda^{A=1}(t)}{\lambda^{A=0}(t)}$ to be the HR in terms of the potential outcomes. As recently discussed (Aalen et al., 2015; Hernán, 2010; Martinussen et al., 2020), $HR^{PO}(t)$ does not admit a natural causal interpretation because it contrasts hazard rates between two non-identical sub-populations of patients, those who would have survived at a given time $t$ when treated ($T^{A=1,C=\infty} \geq t$) and when untreated ($T^{A=0,C=\infty} \geq t$), respectively. Intuitively, while the treated and untreated patients are comparable at time $t = 0$ due to randomization, as time progresses, the two sub-populations may become less and less balanced with respect to covariates other than the treatment. Weaker or more “frail” patients may have survived because they received the treatment, and they would not have survived had they been placed in the untreated group. As a result, the treated are generally more frail than the untreated.

The aforementioned selection bias resembles the issue of truncation by death (Hayden et al., 2005; Rubin, 2006). One solution employs the principal stratification approach (Frangakis & Rubin, 2002) to focus instead on the SACE, which compares the outcomes under different treatment levels among the sub-population who would have survived regardless of the treatment assignment. As recently discussed (Martinussen et al., 2020), an approach similar to the SACE can be used to remedy the selection problem in HR. The causal HR,

$$HR^{C}(t) = \lim_{dt \to 0} \frac{\sum_{i=1}^{n} I[t \leq T^{A_i=1,C=\infty} < t + dt \mid T^{A_i=1,C=\infty} \geq t]}{\sum_{i=1}^{n} I[t \leq T^{A_i=0,C=\infty} < t + dt \mid T^{A_i=0,C=\infty} \geq t]}$$

contrasts the instantaneous risk at time $t$ of the treatment versus no treatment, in the sub-population containing patients who would have survived up to time $t$ regardless of their treatment assignment. Because $HR^{C}(t)$ is a contrast between hazard rates defined on the same sub-population, it is a well-defined causal effect.
Analogous to challenges with identification of the SACE, even had we were able to avoid censoring, for each patient we would have known the survival status at time \( t \) under one treatment value only. Therefore, we do not know which patients would have survived until time \( t \) regardless of their treatment assignment and cannot identify \( HR_C(t) \) from the observed data using standard identification assumptions. Point-identification of the SACE relies on strong assumptions (Hayden et al., 2005; Zehavi & Nevo, 2021) that are unlikely to hold in our setup (Martinussen et al., 2020). Therefore, to avoid making strong and implausible assumptions, we focus on a sensitivity analysis approach.

4 | A SENSITIVITY ANALYSIS APPROACH

When a causal effect is not identified non-parametrically, sensitivity analyses are often resort to. The idea is to identify the causal effect up to an unknown and unidentifiable sensitivity parameter. Then, an estimator of the causal effect is obtained as a function of the unknown sensitivity parameter.

We adopt a frailty point of view to model the unidentifiable cross-world dependence between event times under different treatment values \( T^{A=1, C=\infty} \) and \( T^{A=0, C=\infty} \). Frailty variables are commonly used in the survival analysis literature to model unobserved heterogeneity with known source, for example, to account for and study dependence in clustered survival data (Clayton & Cuzick, 1985; Hougaard, 2000). More recently, a frailty approach was taken to model cross-world dependence between event times under different treatments (Aalen et al., 2015; Martinussen et al., 2020; Nevo & Gorfine, 2022; Stensrud et al., 2017).

We make the following working assumption on the frailty variable \( V \).

Assumption 5. There exists an unmeasured time-fixed frailty variable \( V \), from a known parametric family with mean \( 1 \) and variance \( \theta \), such that

(i) Conditionally on the frailty, the potential event times are independent, \( T^{A=1, C=\infty} \perp T^{A=0, C=\infty} \mid V \).
(ii) Multiplicity assumption, \( \lambda^{A=a}(t) \mid V = v \) = \( \psi^{A=a}(t) \), where \( \psi^{A=a}(t) \) is a function that does not depend on \( V \).

Importantly, we stress that our frailty-based approach is a working sensitivity analysis model. One might argue that a variable like \( V \) leading to independence between the potential survival times does not exist. Here, we use the frailty approach simply to account for the cross-world dependence. In Section 6, we consider in simulations the robustness of our approach to misspecification of the frailty model. The following proposition builds on Assumption 5 to identify \( HR_C(t) \) in a randomized trial setting, as a function of terms estimable from the observed data and of the unknown and unidentifiable variance of the frailty distribution.

**Proposition 1.** Under Assumptions 1–3, and 5, the \( HR_C(t) \) is identified by

\[
HR_C(t) = \frac{\lambda(t|A = 1)}{\lambda(t|A = 0)} \varphi[\Lambda(t|A = 1), \Lambda(t|A = 0), \theta],
\]

where \( \lambda(t|A = a) = \frac{E[dN(t|A = a)]}{E[Y(t|A = a)]} \), \( \Lambda(t|A = a) = \int_0^t \lambda(u|A = a) \, du \), and \( \varphi[\Lambda^{A=1}(t), \Lambda^{A=0}(t), \theta] \) is a function depending on \( \Lambda^{A=1}(t) \), \( \Lambda^{A=0}(t) \) and \( \theta \) and may take a closed form depending on the specific parametric family distribution for \( V \).

The proof is given in Web Appendix A.1.

For observational studies, we replace the randomization assumption (Assumption 3) with conditional exchangeability (Assumption 4) and present an IPTW-based identification formula for \( HR_C(t) \). Let \( \pi(Z) = \text{Pr}(A = 1|Z) \) be the propensity score, namely the probability of being treated given \( Z \), we also modify Assumption 5(ii) such that conditionally on the frailty and \( Z \), the potential event times are independent, \( T^{A=1, C=\infty} \perp T^{A=0, C=\infty} \mid (V, Z) \). Under this modified assumption, our approach does not lead to a closed-form identification formula for \( HR_C(t) \). However, we suggest an identification formula which approximates the true \( HR_C(t) \), that should work well when the event of interest is rare, as is often the case, for example, in cancer epidemiology studies, and/or when the association between \( T^{A=a, C=\infty} \) and \( Z \) is low. We further investigate this issue in Section 6, and illustrate that under a high censoring rate and/or weak association between \( T^{A=a, C=\infty} \) and \( Z \), an estimator built upon our approximated identification formula for \( HR_C(t) \) can be robust even in the case Assumption 5(ii) does not hold exactly.

In Web Appendix A.2, we show that when the event is rare and/or the association between \( T^{A=a, C=\infty} \) and \( Z \) is weak, then under Assumptions 1, 2, 4, and 5, \( HR_C(t) \) can be approximated by

\[
HR_{C\text{IPTW}}(t) = \frac{\lambda_{\text{IPTW}}(t|A = 1)}{\lambda_{\text{IPTW}}(t|A = 0)} \varphi[\Lambda_{\text{IPTW}}(t|A = 1), \Lambda_{\text{IPTW}}(t|A = 0), \theta],
\]

where

\[
\lambda_{\text{IPTW}}(t|A = a) = \frac{E_{\text{IPTW}}[dN(t|A = a) |Z]}{E_{\text{IPTW}}[Y(t|A = a) |Z]}, \quad \Lambda_{\text{IPTW}}(t|A = a) = \int_0^t \lambda_{\text{IPTW}}(u|A = a) \, du.
\]
where $\lambda^{PTE}(t|A = a) = \frac{E[w(Z)\Delta N(t|A = a)]}{E[w(Z)\Delta Y(t|A = a)]}$ and $\Lambda^{PTE}(t|A = a) = \int_0^t \lambda^{PTE}(u|A = a)\,du$, and $\varphi(\cdot)$ is identical to the one appearing on Equation (3).

We can now propose a general sensitivity analysis approach for $HR^C(t)$ based on Equations (3) and (4). The first step is choosing the parametric family distribution for $\lambda$ and a range of possible $\theta$ values. One natural choice for the frailty distribution is the Gamma distribution. In Web Appendix A, we present closed-form formulas for $\varphi$ and the identification of $HR^C(t)$ under different distribution choices. Although the identification formulas depend on the frailty distribution specification, one might expect, based on previous empirical evidence in frailty modeling (Gorfine et al., 2012), that the resulting differences in the $HR^C(t)$ estimator will be small. We demonstrate this point in our analyses of the motivating examples (Section 7). In both examples, the differences in the $HR^C(t)$ estimator under different frailty distributions were mild and did not change the analysis conclusions. The range of $\theta$ values can be selected, for example, by inverting a corresponding range of Kendall’s $\tau$ correlation values between $T^A=0, C=\infty$ and $T^A=1, C=\infty$ (Oakes, 1989). For example, under the Gamma frailty, Kendall’s $\tau$ between $T^A=0, C=\infty$ and $T^A=1, C=\infty$ is $\frac{\theta}{\theta+2}$.

The next step is estimating the identifiable but unknown quantities. Under randomization, these are the functions $\lambda(t|A = a)$ and $\Lambda(t|A = a)$ in Equation (3). In the presence of confounders, we need to estimate $w(Z)$, as well as $\lambda^{PTE}(t|A = a)$ and $\Lambda^{PTE}(t|A = a)$ in Equation (4). In the final step of the sensitivity analysis, the obtained estimators are plugged in Equation (3) or (4) to obtain an estimator for the $HR^C(t)$ function, $\hat{HR}^C(t)$, for a range of $\theta$ values chosen by the researcher.

5  |  ESTIMATION

We present two estimation methods, a semi-parametric Cox-based approach and a non-parametric kernel-based approach. Both estimation methods rely on the independent censoring assumption (Assumption 1). In Section 8, we discuss a potential alternative approach when this assumption is not plausible.

5.1  |  Cox-based estimation

The Cox-based approach assumes that the hazard function of the potential event times $\lambda^{A=a}(t)$ follows the marginal structural Cox PH model (Hernán et al., 2000)

$$\lambda^{A=a}(t) = \lambda_0(t) \exp(\beta a). \quad (5)$$

Since under randomization $\lambda^{A=a}(t) = \lambda(t|A = a)$, then Equation (3) becomes a function of $\lambda_0(t) = \int_0^t \lambda_0(u)\,du$ and $\beta$. For example, in Web Appendix B1, we show that under Gamma frailty, by substituting Equation (5) back into Equation (3), we get that $HR^C(t) = \exp[\beta \exp[\beta \lambda_0(t)\exp(\beta) - 1]]$, and the proposed estimator for $HR^C$ for each value of $\beta$ is

$$\hat{HR}^C(t) = \exp \{ \hat{\beta} \exp[\beta \lambda_0(t)\exp(\hat{\beta}) - 1] \}, \quad (6)$$

where $\hat{\beta}$ and $\hat{\lambda}_0(t)$ are obtained by fitting a standard Cox model.

In the presence of confounders, we propose to estimate $\beta$ and $\lambda_0(t)$ by first estimating the weights $w(Z)$ and then fitting a weighted Cox model (Cole & Hernán, 2004). Estimators for the propensity score $\hat{\pi}(Z)$ can be based on, for example, a logistic regression model, and then the weights are obtained by $\hat{w}(Z) = \frac{\hat{\pi}(A=1)}{\hat{\pi}(A=0)} + \frac{1}{\hat{\pi}(A=0) - \hat{\pi}(A=1)}$. Alternatively, the stabilized weights $\hat{w}^{st}(Z) = \frac{\hat{w}(A=1)}{\hat{\pi}(A=1)} + \frac{\hat{w}(A=0)1 - A}{1 - \hat{\pi}(A=0)}$ can be used to gain reduced variance (Robins et al., 2000).

5.2  |  Kernel-based estimation

As a flexible non-parametric alternative to the Cox-based estimation, we propose to estimate $\lambda(t|A = a)$ or $\lambda^{PTE}(t|A = a)$ at each treatment arm separately without imposing any relationship between hazards under different treatment values (other than the frailty). Furthermore, we do not impose a model for $\lambda^{A=a}(t)$ and specifically we avoid the PH assumption.

5.2.1  |  Pre-smoothed estimators

We first look at a pre-smoothed estimator, which is the building block of the kernel estimator. Under randomization, $\lambda(t|A = a)$ can be estimated by

$$\hat{\lambda}^{pre}(t|A = a) = \frac{1}{n} \sum_{i=1}^{n} \frac{dN_i(t|A_i = a)}{\sum_{j=1}^{n} Y_j(t|A_j = a)}, \quad (7)$$

where $\sum_{i=1}^{n} dN_i(t|A_i = a)$ and $\sum_{j=1}^{n} Y_j(t|A_j = a)$ are the numbers of observed events and of at-risk patients at time $t$ in treatment arm $A = a$, respectively. The cumulative hazard of the observed data distribution, $\Lambda(t|A = a)$, is then estimated at each treatment arm by the well-known Nelson–Aalen estimator $\hat{\Lambda}^{pre}(t|A = a) = \int_0^t \hat{\lambda}^{pre}(u|A = a)\,du$. When confounders are present, the
IPTW estimator analog of (7) is

\[ \hat{\lambda}_{\text{IPTW}}^{\text{pre}}(t|A = a) = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{w}_i(z_i) dN_i(t|A_i = a)}{\sum_{j=1}^{n} \hat{w}_j(z_j) Y_j(t|A_j = a)}. \]  

(8)

Then, \( \hat{\lambda}_{\text{IPTW}}(t|A = a) \) is estimated by the IPTW version of the Nelson–Aalen estimator

\[ \hat{\lambda}_{\text{IPTW}}(t|A = a) = \int_0^t \frac{\hat{\lambda}_{\text{IPTW}}(u|A = a) du}{\sum_{i=1}^{m_a} K_i \left( \frac{t - T_a^{(i)}}{b(t)} \right) d\hat{\lambda}_{\text{IPTW}}(T_a^{(i)}).} \]  

(10)

As before, the local bandwidth \( b(t) \) can be chosen by minimizing the MSE, but modified estimators of the bias and the variance that include the weights are needed. Further technical details about the kernel-based estimation and our proposed MSE estimation for the weighted scenario are given in Web Appendix C.

### 5.3 Asymptotic properties of \( \hat{HR}^C(t) \)

For each time point \( t \), the Cox-based estimator for \( HR^C(t) \) is consistent and asymptotically normal by the continuous mapping theorem and the delta method, since it is a function of \( \hat{\lambda}_0(t) \) and \( \hat{\beta} \), which are consistent and asymptotically normal (Andersen et al., 2012). The kernel-based estimator \( \hat{\lambda}(t|A = a) \) is asymptotically normal (with a rate typically slower than \( \sqrt{n} \) that depends on the smoothness of \( \lambda(t|A = a) \)), and consistent under certain conditions on the local bandwidth \( b(t) \) (Müller & Wang, 1990). Therefore, given results on weighted estimators, one may expect similar results on \( \hat{\lambda}_{\text{IPTW}}(t|A = a) \) (Robins, 1997; Schaubel & Wei, 2011). Finally, the kernel-based estimators for \( HR^C(t) \) are also consistent and asymptotically normal, again by the continuous mapping theorem and the delta method. However, if the local bandwidth, \( b(t) \), is chosen to minimize the MSE, one cannot expect the estimator’s bias to be zero even for a large sample size (Chapter 5.7 in Wasserman, 2006). Thus, if \( b(t) \) is chosen to minimize the MSE, the confidence intervals may not be centered around the true value of \( HR^C(t) \) and their empirical coverage rate is not guaranteed to be close to the desired level.

### 6 SIMULATION STUDIES

We conducted simulation studies to assess and compare the finite-sample performance of the kernel-based and the Cox-based estimators in two settings: (I) under randomization and (II) in the presence of confounders. For each scenario described below, we simulated 1,000 datasets. Technical details are provided in Web Appendix D.

#### 6.1 Data-generating mechanism

Under the randomization setting (I), for each unit we first generated a frailty variable, \( V \), from a Gamma distribution with mean 1 and variance \( \theta \in \{0.2, 0.8, 2.0, 4.6\} \).
These values correspond to Kendall’s $\tau \in \{0.1, 0.3, 0.5, 0.7\}$ between $T^A=0, C=\infty$ and $T^A=1, C=\infty$. Next, we generated the correlated potential event times $T^{A, C}=\infty$ according to a specified hazard rate $\lambda^{A, a}(t|V)$, $a = 0, 1$. We considered three specifications for $\lambda^{A, a}(t|V)$:

Scenario (Ia) $\lambda^{A, a}(t|V) = V \exp[\log(0.5)a + \delta t \exp[\log(0.5)a]]$: This model results in the Cox model (5) for the marginal hazard $\lambda^{A, a}(t)$ with a time-varying HR $C(t)$ (see Martinussen et al. (2018) and Web Appendix D.1.1).

Scenario (Ib) $\lambda^{A, a}(t|V) = V \exp[1.5t + \log(0.5)a]$. Under this model, $\lambda^{A, a}(t)$ does not follow the Cox model (5), and HR $C(t) = 0.5$ for all $t$ (Web Appendix D.1.2).

Scenario (Ic) There are two independent frailties, $V_1 \sim \text{Gamma}(1/\theta, 1/\theta), V_2 \sim \text{Gamma}(\alpha, \gamma)$.

$$\lambda^{A, a}(t|V_1, V_2) = \frac{\alpha}{\gamma} V_1 V_2 \exp[\log(0.5)a + \delta t \exp[\log(0.5)a]] + \frac{V_1}{\delta \alpha} \left[ \exp(\delta t \exp[\log(0.5)a] - 1) \right].$$

Under this model, the potential event times are independent only conditionally on both frailties $V_1, V_2$. In addition, under this model $\lambda^{A, a}(t)$ follows the Cox model (5). See also Web Appendix D.1.3 and Section A.2 in Martinussen et al. (2020).

We generated the censoring time $C^{A=1} = C^{A=0} = C$ from an exponential distribution, whose rate parameter was chosen to yield the desired censoring rate. We considered low to moderate censoring rates ($10\%, 20\%, 40\%$). The actual treatment $A$ was generated with probability $\Pr(A = 1) = 0.5$. The observed time $X$ and the event indicator $\delta$ were determined according to the actual treatment value $A$. We considered sample sizes $n \in \{500, 1000, 5000\}$.

For the observational studies setting (II), we used a data-generating mechanism similar to Setting (I) with the following modifications. For each unit, we simulated a single confounder $Z$ from a standard normal distribution $N(0, 1)$. We simulated the potential event times $T^{A=0, C=\infty}, a = 0, 1$, according to the following hazard rate:

Scenario (II) $\lambda^{A, a}(t|V, Z) = V \gamma \exp[\log(0.5)a + \rho_Z Z]$.

That is, the conditional hazard followed a Cox model. The exact value of $\gamma$ was selected to yield the desired event rates. We considered different $\rho_Z$ values $(\log(0.1), \log(0.5), \log(0.9))$. Under this model, Assumption 5(ii) does not hold, and HR $C(t) = 0.5$ for all $t$ (Web Appendix D.1.4).

For Setting (II), administrative (fixed) censoring was imposed at $t = 10$. Mimicking a large epidemiological cohort, we considered a relatively rare outcome by taking a low event rate ($1\%, 3\%, 5\%, 10\%, 30\%$) and a larger sample size of $n = 50, 000$. The treatment was generated with probability of $\pi(Z) = \frac{\exp[\log(0.5)Z]}{1+\exp[\log(0.5)Z]}$.

6.2 | Analyses

For each scenario, we estimated HR $C(t)$ using both the kernel-based and Cox-based estimators as implemented in our R package CausalHR. In Scenarios (Ia), (Ib), and (Ic), we estimated for each arm the hazard over a grid of 51 time points, taking the minimal time at which the hazard was estimated to be the first event time, and the maximal time to be the time at which 10 patients remained at risk.

In Scenario (II), we first estimated the weights $w(Z)$ via a logistic regression of $A$ on $Z$, before estimating $\text{HR}^C(t)$ using IPTW as described in Section 5. The minimal and maximal times were taken to be $t = 0$ and $t = 10$. We also compared the results to the HR estimated by a conditional Cox model that included the treatment $A$ and the confounder $Z$ as covariates (but not the unobserved frailty).

For all methods and scenarios, standard errors were estimated by the bootstrap with 500 repetitions, and 95% pointwise CIs were calculated by the percentile method.

6.3 | Results

We present here the main results, under Kendall’s $\tau = 0.7$. Additional results under different values of Kendall’s $\tau$, censoring rates, and sample sizes, as well as computation times, are presented in Web Appendix D.

In Scenario (Ia), under which the Cox model (5) was met, both estimation methods performed well (Figure 2). The bias and the empirical standard deviation of the estimates were small for all $t$. When the true values HR $C(t)$ were very close to zero (after $t = 1$), the standard errors were overestimated, especially for the kernel-based estimator. The empirical coverage rates of the 95% CIs for both methods were close to the desired level (Web Figure D.1). As expected, for smaller sample sizes ($n = 500, 1000$) the empirical standard deviations of the estimates were larger and the standard errors of both estimators were overestimated for all $t$ (Web Figure D.2). For higher censoring rates of 40%, the standard errors were still underestimated (Web
FIGURE 2  Performance of the Cox-based and kernel-based estimators in Scenario (Ia) under sample size of $n = 5,000$, 20% censoring rate, and a Gamma frailty distribution with Kendall’s $\tau = 0.7$. Results are presented only until $t = 2$. The two left plots present the mean estimated $HR^C(t)$ across the simulations, plus/minus one empirical standard deviation. The dashed line represents the true $HR^C(t)$. The right plot presents the ratio between the mean estimated standard error (EST.SE) and the empirical standard deviation of the estimates (EST.SE).

$\hat{HR}^C_{Cox}(t)$: Cox-based estimator $\hat{HR}^C_{kernel}(t)$: kernel-based estimator.

FIGURE 3  Performance of the Cox-based and kernel-based estimators in Scenario (Ib) under sample size of $n = 5,000$, 20% censoring rate, and a Gamma frailty distribution with Kendall’s $\tau = 0.7$. Results are presented only until $t = 8$. The two left plots present the mean estimated $HR^C(t)$ across the simulations, plus/minus one empirical standard deviation. The dashed line represents the true $HR^C(t)$. The right plot presents the ratio between the mean estimated standard error (EST.SE) and the empirical standard deviation of the estimates (EST.SE).

$\hat{HR}^C_{Cox}(t)$: Cox-based estimator $\hat{HR}^C_{kernel}(t)$: kernel-based estimator.

Figure D.3), but the bias remained less than 10% for $t \leq 1$. For lower Kendall’s $\tau$ values, the estimators’ performance was generally similar (Web Figures D.4 and D.5).

In Scenario (Ib), when the Cox model (5) was not met, the kernel-based method performed better than the Cox-based method in some aspects, especially in terms of bias (Figure 3). The bias of the kernel-based estimator was smaller for most values of $t$. The empirical standard deviation of the kernel-based estimates increased with $t$ as the available sample size for estimation was decreased. The standard errors were generally well estimated for both methods, although for later time points they were overestimated. In this scenario, there was an under-coverage of the confidence intervals for both methods (Web Figure D.6).

For the kernel-based estimator, this result can be explained by the bias induced by choosing the bandwidth to minimize the MSE (Section 5.3). For the smaller sample sizes ($n = 500, 1000$) the empirical standard deviations of the estimates were bigger and the standard errors were overestimated for all $t$ and for both estimation method (Web Figure D.7). For the higher censoring rate of 40%, the standard errors were larger and overestimated for later time points (Web Figure D.8). For lower Kendall’s $\tau$ values, the bias of the kernel-based estimator was smaller and
the standard errors were generally better estimated (Web Figures D.9 and D.10).

In Scenario (Ic), when Assumption 5 was not met but the marginal model was Equation (5), both estimation methods performed reasonably well with negligible bias (Web Figures D.11–D.14), even though the working frailty model was misspecified.

In Scenario (II) (Figure 4), the bias of the kernel-based method was generally small for nearly all time points, whenever the event rate was small and the confounder effect on the event time was not large. Bias observed at early time points could be explained by low cumulative event rates, while bias at late time points is due to the small number of events. These phenomena are expected for non-parametric estimators applied to survival data. For non-large effects of $Z$ ($\beta_Z = \log(0.5), \log(0.9)$), the relative bias of the kernel-based estimator decreased with the decrease in the event rate. For weak $Z$ effect, the relative bias was small even when the event rate was relatively non-rare (30%). For a very strong confounder effect ($\beta_Z = \log(0.1)$), the relative bias of the kernel-based estimator was approximately 20% regardless of the event rate. Similar trends were observed for the Cox-based estimator, although the relative bias was larger when the event rate was large.

For nearly all event rates and $\beta_Z$ values considered, the relative bias of the proposed estimators was substantially lower than the bias of the conditional Cox estimator. The relative bias of the conditional Cox estimator reached the minimum when the event rate was the smallest and the confounder’s effect was the weakest.

Generally, standard errors were well-estimated for the proposed methods for all $\beta_Z$ values and event rates (Web Figure D.15). The empirical coverage rates of the 95% CIs from the kernel-based approach were close to the desired level for weaker effects of $Z$ ($\beta_Z = \log(0.5), \log(0.9)$). The empirical coverage rates of the 95% CIs from the Cox-based approach were generally far from the desired level, unless the event rate was low (Web Figure D.16). For Kendall’s $\tau = 0.5$, the same trends were observed, but the relative bias of all methods was smaller (Web Figure D.17).

## 7 | REAL-DATA EXAMPLES

We applied the proposed approach to the two real-data examples described in Section 2. At each dataset, we estimated the Cox-based (denoted by $\hat{HR}_{Cox}(t)$) and the kernel-based estimator ($\hat{HR}_{Kernel}(t)$) over a grid of 51 time points. The minimal and maximal time points were chosen in similar way to Section 6. We compared between choosing Gamma and Inverse Gaussian (IG) laws for the frailty distribution. Under the Gamma distribution, we
FIGURE 5 Kernel-based estimation along with plus/minus one standard error for the $HR^C(t)$ in the IMvigor211 study, under different frailty distributions and Kendall’s $\tau$ values. For IG frailty distribution, Kendall’s $\tau$ is bounded by 0.5. The horizontal dashed line represents $HR^C(t) = 1$. For clarity, we present the estimation every 5 time points.

7.1 Effectiveness of Atezo treatment in urothelial carcinoma patients

In this dataset, there was evidence against the validity of the PH assumption ($p < 0.01$). As demonstrated in Section 6 (Scenario (Ib)), when the marginal hazard rate $\lambda^A(t)$ does not follow the Cox model, the Cox-based sensitivity analysis is expected to be biased. Therefore, we focus on our proposed kernel-based analysis. For completeness, the results of the Cox-model sensitivity analysis are presented in Web Appendix E.1. Generally, under all frailty distributions and Kendall’s $\tau$ values, $\hat{HR}^C_{\text{kernel}}(t)$ was larger than one at the start of the study, indicating that Atezo treatment is inferior to Chemo in terms of short-term survival (Figure 5). After approximately four months, $\hat{HR}^C_{\text{kernel}}(t)$ decreased below one, suggesting that Atezo treatment prolongs the survival times of patients with urothelial carcinoma for those who would have survived up to 4 months regardless of their treatment assignment.

In contrast to our findings, the main result in Powles et al. (2018), which conducted the IMvigor211 study, is that “Atezolizumab was not associated with significantly longer overall survival than chemotherapy in patients”. As demonstrated in Section 2, the results of a piecewise Cox PH model (which was not considered by Powles et al. (2018)) would have aligned better with our conclusions.

7.2 Survival time after kidney transplantation

The DIVAT is an ongoing prospective cohort study following kidney transplant recipients. The data include medical records for kidney and/or pancreas transplant recipients from 1990 to 2021 from eight different medical centers in France (Le Borgne et al., 2016). Our analysis included 1281 adults between ages 40 and 70, of which 400 (31%) had
the event of interest, a graft failure, defined as a return to dialysis or death. The event times of the remaining 881 (69%) patients were censored. The treatment variable was donor category, classified into expanded criteria donor (ECD, $A = 1$) and standard criteria donor (SCD, $A = 0$). ECD is defined as donors older than 60, or aged 50–59 with at least two of the following characteristics: history of hypertension, cerebrovascular accident as the cause of death, or terminal serum creatinine higher than 1.5 mg/dl. Of the 1281 patients, 562 (56%) patients received a transplant from ECD donors, and the rest (719, 44%) from SCD donors. Our analysis included the following confounders: recipient’s age, the presence of at least four human leucocyte antigen (HLA) incompatibilities between the donor and the recipient, and whether the transplantation was a re-transplantation.

We fitted a logistic regression for $\pi(Z)$ and verified that the balance was improved after applying the weights (Web Tables F.1 and F.2, and Web Figures F.1–F.3). Those receiving transplant from ECD donors were generally older (OR=1.20, 95% CI: 1.18–1.23). Additionally, second transplantations were more likely for ECD donors (OR=1.29, 95% CI: 0.92–1.83). To improve precision, we used stabilized weights (Section 5.1). To avoid increased standard errors due to large weights, we truncated the weights at the 99th percentile. Density plots of the obtained weights and the propensity scores across the treatment groups are shown in Web Figures F.4 and F.5.

Weighted and non-weighted Kaplan–Meier curves (Cole & Hernán, 2004) corresponding to the two donor categories showed that recipients from SCD donors survive longer (Web Figure F.6, $p < 0.0001$). A naive Cox-model analysis including the donor criterion type and the confounders estimated an HR of 1.59 (95% CI: 1.26–2.01, Web Table F.3) for ECD versus SCD donors. A weighted Cox model estimated an HR of 1.56 (95% CI: 1.19–2.05). No violation of the PH assumption was detected ($p = 0.48$).

Turning to the proposed sensitivity analysis, since there was no evidence against the validity of the PH assumption, we considered both the Cox-based and kernel-based estimators. The Cox-based estimator $\text{HR}^C(t)$ was larger than one and monotonically increased over time (Figure 6). This result suggests that transplantation from ECD donors has a harmful effect on graft failure time, and furthermore, that the impact on the hazard becomes more and more substantial with $t$ among those who would have survived beyond time $t$, regardless of the donor type. The kernel-based method estimated an increasing $\text{HR}^C(t)$ until approximately $t = 5$ years and then either constant or slightly decreasing (but larger than one) $\text{HR}^C(t)$ at the remaining time points.

The estimated coefficients of the confounders were relatively small (Web Table F.3) and the event rate was 31%. In light of the simulation results, we might expect the relative bias of the kernel-based estimator to be lower than those of the Cox-based estimators.

A systematic review of 32 publications on the differences between ECD and SCD found that only five studies reported confounder-adjusted HRs (Querard et al., 2016). Each of these five studies reported a single larger-than-one HR, suggesting that ECD recipients have a poorer prognosis compared to SCD recipients. The conclusions derived from our analysis are similar to these results. However, the estimated $\text{HR}^C(t)$ suggests that the donor category does not have a constant effect on graft survival over time.

8 | DISCUSSION

The HR measure is still one of the most ubiquitous tools for survival analysis in spite of its problematic causal interpretation. Martinussen et al. (2020) has proposed the causal HR, a well-defined alternative causal effect. In this paper, we propose a sensitivity analysis framework for identification coupled with Cox-based or kernel-based estimation of the $\text{HR}^C(t)$. Our simulation results demonstrate that when the marginal structural Cox model is correctly-specified, the Cox-based estimator is preferable over the kernel-based estimator. When the Cox model was misspecified, the kernel-based estimator was less biased, even when the frailty assumption (Assumption 5) was not exactly met. As with many non-parametric methods, a larger sample size might be needed to obtain satisfactory finite-sample properties.

The assumption that the potential survival times are independent given the frailty, is a cross-world independence assumption. This assumption cannot be verified, but as previously mentioned, we adopt it as a working sensitivity analysis framework, and not as a belief in the existence of such a variable $V$. The simulation results (Section 6) indicate that even when this assumption is violated, our approach could still have minimal bias. In many scenarios, a common alternative strategy constructs bounds for causal effects, for example, under the monotonicity assumption (Zhang & Rubin, 2003). Here, however, a monotonicity-like assumption may constrain the causal HR to be equal to zero (Section A.5 in Martinussen et al. (2020)), and therefore we did not follow this strategy.

Our methods were developed under an independent censoring assumption (Assumption 1). In many real-data examples, this assumption is not plausible. One potential
remedy is incorporating Inverse Probability of Censoring Weighting (IPCW) (Robins & Finkelstein, 2000) into the analysis. For observational studies, this means replacing the weights with the product of the IPTW and IPCW weights (Cole & Hernan, 2008; Schaubel & Wei, 2011).

Choosing the bandwidth to minimize the MSE may result in CIs with under-coverage, a well-known phenomenon in non-parametric inference. A potential solution might be under-smoothing (Wasserman, 2006), sacrificing larger standard errors (and wider CIs) to reduce the bias. Another potential extension of the proposed approach is to improve its efficiency by developing an estimation approach that allows for inclusion of more outcome predictors, even in an RCT setting. One potential way to achieve this is use a time-varying coefficient Cox model (Tian et al., 2005; Zucker & Karr, 1990) rather than the
standard Cox model to improve the Cox-based estimator when the PH assumption is unlikely to hold.

ACKNOWLEDGMENTS
The authors would like to thank to anonymous reviewers, the Associate Editor, and the Co-Editor for helpful comments and suggestions. This research was supported by the Israel Science Foundation (ISF grant No. 827/21).

DATA AVAILABILITY STATEMENT
The IMvigor211 data that support the findings in this paper are available from the R package KONPsurv [https://CRAN.R-project.org/package=KONPsurv] (Gorfine et al., 2020). The DIVAT data that support the findings in this paper are available from the R package RISCA [https://CRAN.R-project.org/package=RISCA] (Le Borgne et al., 2016).

ORCID
Rachel Axelrod © https://orcid.org/0000-0002-4721-2591
Daniel Nevo © https://orcid.org/0000-0002-9770-827X

REFERENCES
Aalen, O.O., Cook, R.J. & Raysland, K. (2015) Does Cox analysis of a randomized survival study yield a causal treatment effect? Lifetime Data Analysis, 21(4), 579–593.
Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (2012) Statistical models based on counting processes. Springer Science & Business Media.
Andersen, P.K., Pohar Perme, M., van Houwelingen, H.C., Cook, R.J., Joly, P., Martinussen, T. et al. (2021) Analysis of time-to-event for observational studies: guidance to the use of intensity models. Statistics in Medicine, 40(1), 185–211.
Assel, M., Sjoberg, D., Elders, A., Wang, X., Huo, D., Botchway, A. et al. (2019) Guidelines for reporting of statistics for clinical research in urology. The Journal of Urology, 201(3), 595–604.
Clayton, D. & Cuzick, J. (1985) Multivariate generalizations of the proportional hazards model. Journal of the Royal Statistical Society: Series A, 148(2), 82–108.
Cole, S.R. & Hernán, M.A. (2004) Adjusted survival curves with inverse probability weights. Computer Methods and Programs in Biomedicine, 75(1), 45–49.
Cole, S.R. & Hernán, M.A. (2008) Constructing inverse probability weights for marginal structural models. American Journal of Epidemiology, 168(6), 656–664.
Cox, D.R. (1972) Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202.
Frangakis, C.E. & Rubin, D.B. (2002) Principal stratification in causal inference. Biometrics, 58(1), 21–29.
Gorfine, M., De-Picciotto, R. & Hsu, L. (2012) Conditional and marginal estimates in case–control family data-extensions and sensitivity analyses. Journal of Statistical Computation and Simulation, 82(10), 1449–1470.
Gorfine, M., Schlesinger, M. & Hsu, L. (2020) K-sample omnibus non-proportional hazards tests based on right-censored data. Statistical Methods in Medical Research, 29(10), 2830–2850.
Grambsch, P.M. & Therneau, T.M. (1994) Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3), 515–526.
Hayden, D., Pauler, D.K. & Schoenfeld, D. (2005) An estimator for treatment comparisons among survivors in randomized trials. Biometrics, 61(1), 305–310.
Hernán, M.A. (2010) The hazards of hazard ratios. Epidemiology, 21(1), 13.
Hernán, M.A., Brumback, B. & Robins, J.M. (2000) Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. Epidemiology, 11(5), 561–570.
Hess, K.R., Serachitopol, D.M. & Brown, B.W. (1999) Hazard function estimators: a simulation study. Statistics in Medicine, 18(22), 3075–3088.
Hougaard, P. (2000) Analysis of multivariate survival data, vol. 564. Springer.
Imbens, G.W. & Rubin, D.B. (2015) Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.
Lang, T.A. & Altman, D.G. (2015) Basic statistical reporting for articles published in biomedical journals: the “Statistical Analyses and Methods in the Published Literature” or the SAMPL Guidelines. International Journal of Nursing Studies, 52(1), 5–9.
Le Borgne, F., Giraudaud, B., Querard, A.H., Giral, M. & Foucher, Y. (2016) Comparisons of the performance of different statistical tests for time-to-event analysis with confounding factors: practical illustrations in kidney transplantation. Statistics in Medicine, 35(7), 1103–1116.
Martinussen, T., Vansteelandt, S. & Andersen, P.K. (2020) Subtleties in the interpretation of hazard contrasts. Lifetime Data Analysis, 26(4), 833–855.
Martinussen, T., Vansteelandt, S. & Kragh Andersen, P. (2018) Subtleties in the interpretation of hazard ratios. arXiv 1810.
Martinussen, T., Vansteelandt, S., & Andersen, P. K. (2018) Subtleties in the interpretation of hazard ratios. arXiv:1810.09192.
Müller, H.-G. & Wang, J.-L. (1990) Locally adaptive hazard smoothing. Probability Theory and Related Fields, 85(4), 523–538.
Muller, H.-G. & Wang, J.-L. (1994) Hazard rate estimation under random censoring with varying kernels and bandwidths. Biometrics, 50(1), 61–76.
Nevo, D. & Gorfine, M. (2022) Causal inference for semi-competitive risks data. Biostatistics, 23(4), 1115–1132.
Oakes, D. (1989) Bivariate survival models induced by frailties. Journal of the American Statistical Association, 84(406), 487–493.
Powles, T., Durán, I., Van Der Heijden, M.S., Loriot, Y., Vogelzang, N.J., De Giorgi, U. et al. (2018) Atezolizumab versus chemotherapy in patients with platinum-treated locally advanced or metastatic urothelial carcinoma (imvigor211): a multicentre, open-label, phase 3 randomised controlled trial. The Lancet, 391(10122), 748–757.
Querard, A.-H., Foucher, Y., Combescure, C., Dantan, E., Larmet, D., Lorent, M. et al. (2016) Comparison of survival outcomes between expanded criteria donor and standard criteria donor kidney transplant recipients: a systematic review and meta-analysis. Transplant International, 29(4), 403–415.
Ramlau-Hansen, H. (1983) Smoothing counting process intensities by means of kernel functions. The Annals of Statistics, 11(2), 453–466.
Robins, J.M. (1997) Marginal structural models. Proceedings of the American Statistical Association, pp. 1–10.
Robins, J.M. & Finkelstein, D.M. (2000) Correcting for noncompliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56(3), 779–788.

Robins, J.M., Hernan, M.A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5), 550–560.

Rubin, D.B. (2006) Causal inference through potential outcomes and principal stratification: application to studies with “censoring” due to death. *Statistical Science*, 21(3), 299–309.

Schaubel, D.E. & Wei, G. (2011) Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. *Biometrics*, 67(1), 29–38.

Stensrud, M.J., Valberg, M., Røysland, K. & Aalen, O.O. (2017) Exploring selection bias by causal frailty models. *Epidemiology*, 28(3), 379–386.

Tian, L., Zucker, D. & Wei, L. (2005) On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469), 172–183.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Zehavi, T. & Nevo, D. (2021) A matching framework for truncation by death problems. *arXiv:2110.10186*.

Zhang, J.L. & Rubin, D.B. (2003) Estimation of causal effects via principal stratification when some outcomes are truncated by “death”. *Journal of Educational and Behavioral Statistics*, 28(4), 353–368.

Zucker, D.M. & Karr, A.F. (1990) Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, 18(1), 329–353.

**SUPPORTING INFORMATION**

Web Appendices, Tables, and Figures referenced in Sections 4–7 are available with this paper at the Biometrics website on Wiley Online Library. The *R* package CausalHR implements our proposed methodology. The package and fully reproducible simulations and data analyses are available at [https://github.com/xlrod1](https://github.com/xlrod1). In addition, a zip file containing all code to replicate the simulation studies and real-data examples is available with this paper at the Biometrics website on Wiley Online Library.

**How to cite this article**: Axelrod, R. & Nevo, D. (2023) A sensitivity analysis approach for the causal hazard ratio in randomized and observational studies. *Biometrics*, 79, 2743–2756. https://doi.org/10.1111/biom.13797