Article

CLINICAL TRIALS

Clinical Trials
1–13
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17407745241243308
journals.sagepub.com/home/ctj
Sage

Causal interpretation of the hazard ratio in randomized clinical trials

Michael P Fay[1] and Fan Li[2]

Abstract

Background: Although the hazard ratio has no straightforward causal interpretation, clinical trialists commonly use it as a measure of treatment effect.

Methods: We review the definition and examples of causal estimands. We discuss the causal interpretation of the hazard ratio from a two-arm randomized clinical trial, and the implications of proportional hazards assumptions in the context of potential outcomes. We illustrate the application of these concepts in a synthetic model and in a model of the time-varying effects of COVID-19 vaccination.

Results: We define causal estimands as having either an *individual-level* or *population-level* interpretation. Difference-in-expectation estimands are both individual-level and population-level estimands, whereas without strong untestable assumptions the causal rate ratio and hazard ratio have only population-level interpretations. We caution users against making an incorrect individual-level interpretation, emphasizing that in general a hazard ratio does not on average change each individual's hazard by a factor. We discuss a potentially valid interpretation of the constant hazard ratio as a population-level causal effect under the proportional hazards assumption.

Conclusion**:** We conclude that the population-level hazard ratio remains a useful estimand, but one must interpret it with appropriate attention to the underlying causal model. This is especially important for interpreting hazard ratios over time.

Keywords

Causal inference, clinical trial, estimand, hazard ratio, proportional hazards model, survival analysis

Introduction

Randomization is an essential tool of clinical research thanks to its three salient properties:

1. Randomization *balances covariates*—observed and unobserved—and therefore guarantees that test and control arms will be similar, with high probability, with respect to potential confounders of the treatment–outcome relationship.
2. Randomization confers a patina of *objectivity*; both scientific and lay audiences acknowledge that a randomized trial offers the most fair and reliable form of evidence about a treatment comparison.
3. Randomization enables *causal inferences* about treatment effects, in that one can answer causal questions without recourse to unverifiable assumptions.

In contemporary parlance, a *causal model* is a statistical model that describes the distributions of the outcomes that we would observe under alternative assignments of individuals to treatment arms. A *causal inference* thus refers to an estimate, test, or interval regarding the parameters of a such a model.

A trialist who seeks to generate a causal inference should define the causal estimand, or the parameter of the distribution in which scientific interest lies.[1] While one can undertake causal inference based on a test alone (for example, the logrank test), it is generally advantageous to estimate a parameter that describes the magnitude and direction of the treatment effect.

[1] Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD, USA
[2] Department of Biostatistics and Center for Methods in Implementation and Prevention Science, Yale School of Public Health, New Haven, CT, USA

Corresponding author:
Michael P Fay, Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane, Room 4c40, Rockville, MD 20852, USA.
Email: mfay@niaid.nih.gov

Such considerations lead to models based on the *potential outcomes* framework, in which each individual has a pair of potential outcomes—the outcome experienced if randomized to test and the outcome experienced if randomized to control. One defines causal estimands with reference to these potential outcomes; two typical examples are

- The average difference between the potential outcomes the trial participants would have had on the test therapy and the outcomes they would have had on control;
- With binary outcomes, the ratio of the probability of a positive potential outcome on the test therapy to the probability of a positive potential outcome on the control.

With uncensored outcomes, one can estimate such parameters using methods that are similar to two-sample procedures from classical statistical inference.[2] With censored survival outcomes, as arise in many clinical trials, it is conventional to assume that the distributions of the observed time-to-event outcomes follow a specified parametric or semiparametric form indexed by a treatment effect parameter; one then estimates and interprets the treatment effect parameter. The Cox[3] proportional hazards model is in common use for this purpose because it can describe a range of data-generating processes without requiring full distributional assumptions. Because the key parameter of the Cox model is the *hazard ratio*, defining the ratio of instantaneous risk of the event between the two arms, it is tempting to base causal inferences on this parameter in clinical trials with censored survival outcomes.

Yet as others have shown, and we explain below, causal interpretation of the hazard ratio requires some care.[4–7] Thus, a clinical trialist who wishes to base inferences on censored survival data must exercise caution in interpreting the hazard ratio, even when analyzing a study that enjoys the protective effect of randomization.

We begin by reviewing causality in light of the models and parameters that we commonly employ to analyze two-arm randomized trials. After illustrating concepts with some simple examples, we turn to trials with right-censored survival outcomes, describing estimands and interpretation under both proportional and nonproportional hazards assumptions. Finally, we illustrate our ideas in two numerical data examples. Our focus throughout is on the identifiability and interpretation of estimands, with only passing mentioning of estimators.

Causal inference in a randomized clinical trial

Potential outcomes

Contemporary causal inference often relies on the concept of potential outcomes, an idea attributed initially to R. A. Fisher and Jerzy Neyman, and later substantially developed by Donald Rubin[2] and James Robins.[8] This approach posits that an individual has a potential value of the outcome variable corresponding to each treatment or exposure.

For a two-arm randomized trial, an individual's potential outcomes are the values of the outcome under the control and test conditions. We denote the random arm assignment for individual i to be Z_i, where $Z_i = 0$ for control and $Z_i = 1$ for test. The potential outcome pair for individual i is $\{Y_i(0), Y_i(1)\}$, where that individual manifests $Y_i(0)$ if assigned to control and $Y_i(1)$ if assigned to test. Under the Stable Unit Treatment Value Assumption (i.e. treatments are defined unambiguously with no interference between units), the observable outcome is $Y_i = Y_i(Z_i) = Z_i Y_i(1) + (1 - Z_i)Y_i(0)$.

Defining comparison sets

Following Frangakis and Rubin,[9] we consider causal estimands to be those that compare two sets of potential outcomes

$$\{Y_i(0) : i \in \Omega\}\text{and}\{Y_i(1) : i \in \Omega\}, \qquad (1)$$

where Ω is some defined group of individuals. Most commonly, Ω refers to the cohort of trial participants, the population of relevant patients, or some demographic subgroup of the population.

The set Ω need not be finite. In clinical trials, we are typically interested in the average effect of the test therapy as applied to the notional (and infinitely large) *superpopulation* of potential users who satisfy trial entry criteria and from which the study sample constitutes an assumed random sample. The notion of an assumed random sample is an idealized one, as the recruitment process in a specific clinical trial rarely stipulates formal random sampling from a given population. Our technical discussion will refer to this idealized model.

In the superpopulation model, the potential outcomes are assumed to be stochastic and follow a bivariate distribution: $\{Y_i(0), Y_i(1)\} \sim F$. Because we can observe only one potential outcome per individual, without strong assumptions, we can identify only estimands that depend on the marginal distributions for the two arms, $Y_i(0) \sim F_0$ and $Y_i(1) \sim F_1$.

As indicated above, Ω can refer to a set defined by baseline variables such as sex, race, or disease stage. Inference for estimands with such restrictions are straightforward; the fact of randomization enables us to create an unbiased estimate of the causal effect.

When there is a post-treatment intercurrent event such as treatment noncompliance or death, one can define causal estimands for (latent) subsets of individuals such as those who would comply with whatever treatment they are offered[10] or those who would always survive until the final outcome measurement time.[11]

These subsets are denoted *principal strata*,[9] and the associated estimands are referred to as the complier average causal effect and the survivor average causal effect, respectively.[11] Although causal inference is possible in such cases, it generally requires assumptions that are unverifiable from the data.[2]

In what follows, we focus on causal estimands that compare two sets of potential outcomes both from the same clearly defined superpopulation Ω (as in equation (1)), where the identifiable estimands are functionals of F_0 and F_1. Our approach for comparing those sets is driven by the context. Before tackling hazard ratios, we consider those choices for some simpler causal estimands.

Compare, summarize, and interpret

A complete formulation of the causal estimand requires a method of summarization (of the potential outcomes of the individuals in the population) and a method of comparison (of the summaries). We first consider the simple case of uncensored potential outcomes without covariates.

The simplest comparison is the difference $Y_i(1) - Y_i(0)$, which measures how much the test treatment increases the response compared to the control in individual i. We cannot identify any specific individual's causal effect, but under randomization, we can readily identify and estimate the *average treatment effect*,

$$\Delta_{\mathrm{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)]. \tag{2}$$

we refer to Δ_{ATE} as an *individual-level* estimand because it represents a summarization of causal effects defined at the individual level. We emphasize that Δ_{ATE} represents the mean but does not quantify the variation of the individualized treatment effects, that is, the $Y_i(1) - Y_i(0)$. If patients respond to the test condition differently, the individualized treatment effects may vary substantially about Δ_{ATE}.

Alternatively, we can define causal estimands by reversing the order of comparison and summarization. That is, we can summarize each of the two sets $\{Y_i(0) : i \in \Omega\}$ and $\{Y_i(1) : i \in \Omega\}$ first, then compare the summaries. We call that a *population-level* estimand.

The quantity Δ_{ATE} is both an individual-level and a population-level estimand, because it has two interpretations: As an individual-level estimand, it is the average of the differences in potential outcomes on the individuals in Ω (i.e. $\mathbb{E}[Y_i(1) - Y_i(0)]$), while as a population-level estimand, it is the difference between the mean response on test and the mean response on control (i.e. $\mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$). This classification is similar to subject-specific versus population-averaged estimands,[12,13] and just as in that classification some estimands are of one type or the other but not both.

Next consider the case where potential outcomes are positive and the comparison operator is the ratio. If we summarize by expectation, two possible estimands are,

$$\mathbb{E}_F\left[\frac{Y_i(1)}{Y_i(0)}\right] \neq \frac{\mathbb{E}_{F_1}\left[Y_j(1)\right]}{\mathbb{E}_{F_0}\left[Y_i(0)\right]} \equiv \rho. \tag{3}$$

On the left-hand side (using $[Y_i(0), Y_i(1)] \sim F$), we have an individual-level estimand where we compare then summarize, whereas on the right-hand side, we have a population-level estimand where we summarize then compare. Although we may be more interested in the individual-level estimand, the average fold-change caused by the treatment (i.e. $\mathbb{E}_F[Y_i(1)/Y_i(0)]$) depends on the correlation between $Y_i(1)$ and $Y_i(0)$ and is unidentifiable with only one potential outcome observed per participant. Thus, we can identify only the population-level estimand ρ.

Importantly, a value of $\rho = 2$ does not imply that on average, the treatment causes the potential outcome on test for each participant to be double that of their potential outcome on control (i.e. $\rho = 2$ does not imply that $\mathbb{E}_F[Y_i(1)/Y_i(0)] = 2$). Rather it implies that the average outcome had all participants received test is double the average outcome had all participants received control. The estimand ρ is identifiable even when the potential outcomes allow zero values (as with binary data). When $Y_i(z)$ is binary, ρ is the *causal risk ratio*.

If we are interested in the average individual-level ratio effect, we can use the geometric mean to summarize positive potential outcomes. Then we have

$$\exp\{\mathbb{E}_F \log[Y_i(1)/Y_i(0)]\},$$

which is identifiable because it can be re-expressed as

$$\exp\{\mathbb{E}_{F_1} \log(Y_i(1)) - \mathbb{E}_{F_0} \log(Y_i(0))\} = \gamma_1/\gamma_0, \tag{4}$$

where γ_z is the geometric mean of $Y_i(z)$. In other words, for γ_1/γ_0 the order of application of the comparison and summarization is irrelevant. Thus, we can interpret γ_1/γ_0 as the geometric mean of the individual-level ratio effects or as the ratio of geometric mean effects (i.e. the geometric mean in the test arm over the geometric mean of the control arm).

Finally, consider ordinal or continuous potential outcomes and let the comparison operator be a relative effect,

$$g(Y_i(0), Y_i(1)) = I(Y_i(1) > Y_i(0)) + \frac{I(Y_i(1) = Y_i(0))}{2},$$

where the indicator function $I(A)=1$ when A is true and 0 otherwise. Then

$$\begin{aligned}\psi &\equiv \mathbb{E}_F[g(Y_i(0),Y_i(1))] \\ &\neq \mathbb{E}_{F_0}\mathbb{E}_{F_1}\left[g(Y_i(0),Y_j(1))\right] \\ &\equiv \phi.\end{aligned} \tag{5}$$

If larger potential outcomes are better, then the individual-level estimand, ψ, represents the proportion of the study population that would do better on test (corrected for ties), whereas the population-level estimand, ϕ, represents the probability that a randomly selected individual who received test would have a better outcome than a randomly selected individual who received control (again with a correction for ties). The estimand ϕ, called the Mann–Whitney parameter, the relative effect, or the probabilistic index, is the expectation of the Mann–Whitney test statistic, whereas ψ is identifiable from a randomized trial only under strong assumptions. Both ϕ and ψ are causal estimands, but they differ and one must not confuse them.[14,15]

Regardless of the order of comparison and summarization, an important feature of a causal estimand is the restriction to a common Ω (e.g. a common superpopulation). While this requirement is often trivial in the aforementioned examples, complications arise with censored survival data, which we discuss next.

Causal inference with censored survival data

With right-censored survival outcomes, causal estimands based on the mean are undesirable because one cannot directly estimate the mean without modeling assumptions to which inferences may lack robustness. Moreover, nonparametric estimands like the Mann–Whitney parameter ϕ are unsatisfactory because they are sensitive to the censoring distribution, which may vary across studies.

Estimands based on survival curves

If the censoring is independent of the potential outcomes, we can estimate the distribution for each arm up to the end of follow-up, commonly using a Kaplan–Meier estimator.[16] A problem in defining the estimand is that there is some time τ beyond which we have no follow-up data on any individual. One way to address this problem is to prespecify a time after randomization, say $t^* \in (0,\tau]$, and define our estimand with respect to that time. For example, we could focus on the treatment effect contrast at 1 year after randomization ($t^*=1$), and define the estimand as the survival difference,

$$\text{causal survival difference}(t^*) = S_1(t^*) - S_0(t^*),$$

where $S_z(t)=\Pr[Y_i(z)>t]$ is the survival function for the potential outcome $Y_i(z)$. Because $S_z(t)=\mathbb{E}\{I(Y_i(z)>t)\}$, the causal survival difference has both an individual-level (i.e. $\mathbb{E}\{I(Y_i(1)>t^*)-I(Y_i(0)>t^*)\}$) and a population-level (i.e. $S_1(t^*)-S_0(t^*)$) interpretation.

Causal survival ratios or odds ratios (or alternatively, causal failure ratios) can be defined in the same fashion as those for non-censored binary outcomes. As with the binary outcomes, interpretation can be subtle. With censoring, one can use the Kaplan–Meier estimates for each arm, basing inferences on asymptotic[17] or non-asymptotic[18] methods. One can address covariate-dependent censoring through inverse-probability-of-censoring weighting.[19,20]

A closely-related causal estimand is the difference in restricted mean potential survival times (RMST),[21,22]

$$\text{difference in RMST}(t^*) = \int_0^{t^*} S_1(t)dt - \int_0^{t^*} S_0(t)dt,$$

where t^* is a pre-specified restriction time. This is both an individual-level and a population-level estimand.

Inference with right-censored data can also be based on the *accelerated failure time* model.[23] Here

$$S_1(t) = S_0(t\exp(-\beta));$$

that is, the failure time in group 1 is accelerated (or decelerated) by the factor $\exp(-\beta)$, where β is denoted the scale-change parameter. Because time-to-event potential outcomes are positive, a multiplicative treatment effect on time can be expressed as a ratio as in equation (4); therefore, $\exp(-\beta)$ has both an individual-level and population-level interpretation.

Why hazard ratios are difficult to interpret causally

The hazard ratio is a popular measure of treatment effect in event-based trials. Recent scholarship, however, has noted that the usual definition of hazard ratio at each follow-up time t does not possess a straightforward causal interpretation.[4,5,7]

For a two-arm randomized trial consider first the comparison set (equation (1)) where Ω is the superpopulation of individuals who would participate in the study. In the population-level estimand, we summarize $Y_i(z)$ into a hazard function at t, say $\lambda_z(t)$, for each $z=0,1$:

$$\begin{aligned}\lambda_z(t) &= \lim_{h\to 0}\Pr\left[t \leqslant Y_i(z) < t+h \,|\, Y_i(z) \geqslant t\right]/h \\ &= \frac{\lim_{h\to 0}\Pr[t \leqslant Y_i(z) < t+h]/h}{\Pr[Y_i(z) \geqslant t]} \\ &= f_z(t)/S_z(t),\end{aligned}$$

where $f_z(t) = \frac{\partial F_z(t)}{\partial t}$ is the density function of $Y(z)$.

Three elements complicate causal interpretation of the hazard ratio. First, $\lambda_z(t)$ is a derivative—a mathematical nuisance that one can skirt by working with a discrete hazard that approaches the usual (continuous) hazard as the discretization becomes finer. Second, both $\lambda_z(t)$ and $f_z(t)$ are local effects that convey little information about times other than t. Contrast that with $S_z(t)$ which summarizes effects *up to t* (i.e. the experience of members of the set $\{Y_i(z) : i \in \Omega\}$ from time 0 through time t). Third, $\lambda_z(t)$ is either the ratio of two functions, $f_z(t)/S_z(t)$, or alternatively the limit of a scaled conditional probability. As the ratio of two functions, it is not identified with only an expectation of a function using only one random variable $Y_i(z)$—we need two such expectations. As a limit of a scaled conditional probability, we can think of identifying the probability from the conditional set $\Omega_z(t) = \{i : Y_i(z) \geqslant t \text{ and } i \in \Omega\}$.

The key problem for interpreting the hazard ratio at time t, say $\theta(t) = \lambda_1(t)/\lambda_0(t)$, is that if there is a treatment effect, then it is likely that $\Omega_0(t) \neq \Omega_1(t)$, and in that case, $\theta(t)$ does not meet the identical comparison set condition of expression (1). Hernán[4] refers to this as a built-in selection bias arising from the "differential depletion of susceptibles" when there is a treatment effect. Thus, we cannot interpret $\theta(t)$ causally as a hazard ratio in the sense that it represents the ratio of instantaneous risk at t of the same population under different treatment conditions. In another sense, expression (1) is met for $\theta(t)$ because each hazard is a function of the associated marginal distribution, $\lambda_z(t) = \left(\frac{\partial F_z(t)}{\partial t}\right)/(1 - F_z(t))$, and the marginal distributions F_0 and F_1 are identifiable and represent the same comparison set by randomization. Importantly, in either sense, we cannot interpret $\theta(t)$ as an individual-level causal estimand without strong assumptions (see e.g. Supplementary Material Appendix A). For these reasons, interpretations of the hazard ratio as a causal parameter can lead to incorrect conclusions about whether, to what extent, and in what range of times, a treatment is beneficial.[24] An exception occurs when we can assume proportional hazards, as we explain below, but even in this case, the hazard ratio does not describe the causal effects in a mechanistic data-generating manner.

The hazard ratio under proportional hazards

When there are no baseline covariates, the proportional hazards assumption holds when $\lambda_1(t)/\lambda_0(t) = \theta$ for all t, which implies $S_1(t) = S_0(t)^\theta$ for all t. Thus for any time t,

$$\frac{\log S_1(t)}{\log S_0(t)} = \theta. \tag{6}$$

Because $S_z(t)$ summarizes the survival experience for $Y_i(z)$ from 0 to t, equation (6) shows that θ is a population-level causal estimand that compares the summarized experience of the two types of potential outcomes up to t over the superpopulation Ω. This removes all three difficulties of interpretation of the hazard ratio: there are neither limits nor conditional probabilities, and the effects are not local.

Nevertheless, even when $\lambda_1(t)/\lambda_0(t) = \theta$ for all t, we caution that it is a mistake to interpret it as an individual-level causal estimand. That is, an assumption of proportional hazards with a hazard ratio of θ does *not* imply that each individual has their own specific pair of hazard functions (e.g. individual i has the pair λ_{1i} and λ_{0i}) and that the average of the individual specific hazard ratios is θ for all t. Rather, each arm is associated with a hazard function, and the hazard for the test arm is θ times the hazard for the control arm at every t.

Representations of the hazard ratio

To explore other parameterizations of treatment effects on survival, we list here several alternative representations of the dependence of event probabilities on time. In later sections, we explore the causal interpretation of some of those functions without the proportional hazards assumption.

Hazard ratio:

$$\theta(t) = \lambda_1(t)/\lambda_0(t).$$

Cumulative hazard ratio:

$$\theta_{\text{CHR}}(t) = \frac{\Lambda_1(t)}{\Lambda_0(t)} = \frac{\int_0^t \lambda_1(u)du}{\int_0^t \lambda_0(u)du} = \frac{\log S_1(t)}{\log S_0(t)}.$$

Simple average hazard ratio:

$$\theta_{\text{sAHR}}(t) = \frac{\int_0^t \theta(u)w(u)f(u)du}{\int_0^t w(u)f(u)du},$$

where $w(t)$ is a positive, integrable function; $f(t) = (f_0(t) + f_1(t))/2$, and $f_z(t)$ is the density function for $Y_i(z)$.[25] For the two-arm trial example, Prentice and Aragaki[26] discuss using $w(t) = f_0(u)/f(u)$.

Geometric average hazard ratio:

$$\theta_{\text{gAHR}}(t) = \exp\left\{\frac{\int_0^t \log(\theta(u))w(u)f(u)du}{\int_0^t w(u)f(u)du}\right\}.$$

Weighted partial likelihood estimation of the assumed constant hazard ratio parameter (in a model adjusting only for the treatment indicator) corresponds to estimating a geometric average hazard ratio with a specific weight function $w(t)$.[27]

Kalbfleisch and Prentice[28] defined the average hazard ratio estimand as

$$\theta_{\mathrm{AHR}}(t) = \frac{\int_0^t (\lambda_1(u)/\lambda_+(u))w(u)f(u)du}{\int_0^t (\lambda_0(u)/\lambda_+(u))w(u)f(u)du},$$

where $\lambda_+(t) = \lambda_0(t) + \lambda_1(t)$. Different weights can lead to other representations. First, in the extreme case where the weight function $w(t) = \lambda_+(t)/f(t)$, $\theta_{\mathrm{AHR}}(t) = \theta_{\mathrm{CHR}}(t)$. Second, Schemper et al.[25] showed that when the weight function

$$w(t) = \frac{S_0(t)f_1(t) + S_1(t)f_0(t)}{f_1(t) + f_0(t)},$$

then $\theta_{\mathrm{AHR}}(t) = \theta_{\mathrm{LR}}(t)$ and $\lim_{t\to\infty} \theta_{\mathrm{AHR}}(t) = \theta_{\mathrm{LR}}$, where θ_{LR} and $\theta_{\mathrm{LR}}(t)$ are defined next.

Loss ratio:

$$\theta_{\mathrm{LR}} = \frac{\Pr[Y_i(1) < Y_j(0)]}{\Pr[Y_i(1) > Y_j(0)]} = \frac{\int_0^\infty S_0(t)f_1(t)dt}{\int_0^\infty S_1(t)f_0(t)dt},$$

where the i and j subscripts indicate that $Y_i(1)$ and $Y_j(0)$ come from different individuals, and $\theta_{\mathrm{LR}}^{-1}$ is called the *win ratio*. The loss ratio is also called the *odds of concordance*.[25] We can also consider a type of restricted loss ratio (i.e. by evaluating the integral only up to some time t),[29]

$$\theta_{\mathrm{LR}}(t) = \frac{\int_0^t S_0(u)f_1(u)du}{\int_0^t S_1(u)f_0(u)du}.$$

Cox model average hazard ratio:

$$\theta_{\mathrm{CoxAHR}}(t) = \lim_{\min(n_0, n_1)\to\infty} \mathbb{E}_{F_1}\mathbb{E}_{F_0}\left[\hat{\theta}_{\mathrm{mple}}(t)\right], \qquad (7)$$

where $\hat{\theta}_{\mathrm{mple}}(t)$ is the maximum partial likelihood estimator of the hazard ratio from a randomized trial with n_0 and n_1 individuals in the two arms, using only the treatment indicator in the Cox model, and assuming no censoring before t and all outcomes after t are censored at t (here we use t to indicate an administrative censoring time, and therefore, $\hat{\theta}_{\mathrm{mple}}(t)$ is not to be confused with a time-varying hazard ratio estimator). This is the parameter consistently estimated by the simple Cox model from a study up until time $t = \tau$, with only a treatment term and no time-varying effects. Under proportional hazards, we can allow independent right censoring before t, and the Cox estimator is consistent for $\theta_{\mathrm{CoxAHR}}(t)$.

Under proportional hazards $\theta(t) = \theta$ and any of the above estimands has a causal interpretation because they are all equal to θ, and θ is a population-level causal estimand that can be written as a contrast of log survival functions at any time t.

Even under proportional hazards, however, none of these ratios is an individual-level estimand (see Supplementary Material Appendix A for a possible definition of this individual-level estimand). For example, when the Cox model (with treatment included as the only predictor and a time-constant treatment effect parameter) gives an estimate of 0.5, this does not imply that the test therapy reduces each individual's hazard by half, unless unrealistic and untestable assumptions are made (e.g. the assumptions of Supplementary Material Appendix A). Rather, in a population of individuals receiving treatment the hazard rate at any time t is half what it would be in the same population receiving control.

Interpretation under non-proportional hazards

When hazards are non-proportional, the parameters enumerated above represent different parameters. All remain identifiable from a two-arm randomized trial even without the proportional hazards assumption because they are functions of F_1 and F_0. A fundamental difference between $\theta(t)$ and the other parameters is that $\theta(t)$ measures a local effect, the instantaneous hazard ratio at t, whereas the others summarize effects through the time frame from 0 to t.

The loss ratio θ_{LR} is defined without reference to t, and without proportional hazards (or some other assumption), we cannot identify it for the vast majority of studies that include censored observations. Without further assumptions on F_0 and F_1, all of the other parameters depend on t.

Nor does $\theta_{\mathrm{LR}}(t)$, the loss ratio up to time t, have a straightforward causal interpretation. Writing it in terms of conditional probabilities as

$$\theta_{\mathrm{LR}}(t) = \frac{\Pr\left[Y_i(1) < Y_j(0) | Y_i(1) \leq t\right]}{\Pr\left[Y_i(1) > Y_j(0) | Y_j(0) \leq t\right]}, \qquad (8)$$

we see that the conditioning sets do not represent the same individuals, nor do they select the same proportion of individuals in each arm of a 1:1 randomized trial when the treatment has an effect. Thus, this parameter is subject to the same critique as $\theta(t)$.

With the popularity of the Cox model and the availability of software, trialists often use $\theta_{\mathrm{CoxAHR}}(t)$ (with $t = \tau$) to represent the treatment effect even when the proportional hazards assumption fails. Lin and Wei[30] developed a robust variance formula for the Cox estimator that enables asymptotically valid inferences on $\theta_{\mathrm{CoxAHR}}(t)$. Nevertheless, the interpretational challenge is apparent from the complexity of its definition in equation (7). An added difficulty is that under independent right censoring with non-proportional hazards, the estimand depends on the censoring distribution,[31–33] although there is a weighted version that is consistent for an estimand that does not depend on the censoring.[31]

The average hazard ratio estimands—$\theta_{\mathrm{sAHR}}(t)$, $\theta_{\mathrm{gAHR}}(t)$, and $\theta_{\mathrm{AHR}}(t)$—lack easy causal interpretations because of the conditioning issue, although some choices of weights can lead to simplifications. For example, $\theta_{\mathrm{AHR}}(t)$ with weight function $w(t) = \lambda_+(t)/f(t)$ leads to

$\theta_{\mathrm{AHR}}(t) = \theta_{\mathrm{CHR}}(t)$. The most straightforward comparison causal estimand (of those that give the hazard ratio under proportional hazards) appears to be $\theta_{\mathrm{CHR}}(t)$, which remains a population-level causal estimand derived as a contrast of two survival functions; Wei and Schaubel[34] also recommended this causal estimand with non-proportional hazards. More recently, Vansteelandt et al.[35] proposed causal estimands that build on $\log\{\theta_{\mathrm{CHR}}(t)\}$, but adjusting for baseline covariates.

The causal hazard ratio

We have emphasized that $\theta(t)$ does not have a straightforward causal interpretation. Nevertheless, individuals receiving a treatment are interested in the hazard ratio as a potential source of information on how the treatment will affect them. One way to handle this is to define a new type of hazard ratio based on comparisons of potential outcomes within the same conditioning set. For example, Martinussen et al.[24] introduced the *causal hazard ratio*, which changes the conditioning set to include those who would have survived to t regardless of the test and control condition. For example, the *marginal* causal hazard ratio is $\theta_{\mathrm{mcHR}}(t) = \lambda_1^*(t)/\lambda_0^*(t)$, where

$$\lambda_z^*(t) = \lim_{h \to 0} \frac{\Pr[t \leqslant Y_i(z) < t + h | Y_i(0) \geqslant t, Y_i(1) \geqslant t]}{h}.$$

By assigning a common conditioning set in the numerator and denominator of the ratio, equation (1) applies and $\theta_{\mathrm{mcHR}}(t)$ is a population-level causal estimand. Unfortunately, it is not identifiable without further assumptions because $\{Y_i(0), Y_i(1)\}$ are never jointly observed. One such additional (and unrealistic) assumption is independence of $Y_i(0)$ and $Y_i(1)$. If that holds, then $\theta_{\mathrm{mcHR}}(t)$ reduces to $\theta(t)$. Of course, as one never observes both potential outcomes together, there is no way to assess their independence.

Another possibility is to assume a frailty model with known parameter values governing the joint density of $\{Y_i(0), Y_i(1)\}$, in which case, one can identify the causal hazard ratio via a sensitivity analysis.[36] To get a cumulative effect instead of a local effect, one can also construct average causal hazard ratio estimands by replacing $\lambda_z(t)$ with $\lambda_z^*(t)$, in for example $\theta_{\mathrm{sAHR}}(t)$. Because $\theta_{\mathrm{mcHR}}(t)$ and its cumulative average versions are not identifiable without strong assumptions, they may be of primarily theoretical interest.

Baseline covariate adjustment

Below, we discuss two issues with baseline covariate adjustment.

Adjustment to improve efficiency

One can generally improve the efficiency of estimates of Δ_{ATE} given in equation (2) by adjusting for baseline covariates that are predictive of the outcome. Letting X_i be a vector of baseline covariates for individual i, we can rewrite Δ_{ATE} as $\Delta_{\mathrm{ATE}} = \mathbb{E}_X\left[\mathbb{E}_{Y|X}\{Y_i(1) - Y_i(0)|X_i = X\}\right]$, which suggests a form of regression adjustment. In a randomized trial, adjusting for baseline covariates can lead to a smaller asymptotic variance for estimating the average treatment effect. Tsiatis et al.[37] and Wang et al.[38] discussed model-robust approaches to adjust for covariates through analysis of covariance models and semiparametric efficient estimators. Zeng et al.[39] provided the theory of model-robust propensity score weighting estimators for Δ_{ATE}.

Now consider censored data and modeling the survival functions either parametrically (e.g. by a Weibull model) or semiparametrically (e.g. by a proportional hazards or proportional odds model) conditional on treatment and baseline covariates X_i. One can then create individual-specific survival functions and average them over the trial population using a g-formula representation,

$$\Pr[Y_i > t | Z_i = z] = \mathbb{E}_X\{Pr[Y_i > t | Z_i = z, X_i = x]\}, \qquad (9)$$

where $Y_i = Y_i(Z_i)$. This form of covariate adjustment can lead to a more efficient estimator for the causal survival difference or the difference in restricted mean survival.[40]

Although equation (9) applies to any survival function and distribution of covariates, it does not imply that both $\Pr[Y_i > t | Z_i = z]$ and $\Pr[Y_i > t | Z_i = z, X_i = X]$ will come from the same class of models. We consider this issue in the next section.

Hazard ratio with adjustment

A practical way to get hazard ratio estimands that are closer to individual-level estimands is to include baseline covariates in a model. There are many ways to incorporate baseline variables;[41] we focus on the simple method of including them as main effects in a Cox model.

Previously, we considered proportional hazards models with only the treatment indicator Z_i, where

$$\lambda_{Z_i}(t) = \lambda_0(t) \exp(Z_i \beta), \qquad (10)$$

Now suppose the data follow a proportional hazards model when we include baseline covariates in addition to the treatment indicator. Let X_i be a p-vector of baseline covariates for individual i. Assume the hazard function is

$$\lambda^{\star}_{Z_i}(t;X_i) = \lambda^{\star}(t)\exp\big(Z_i\beta^{\star} + X_i\gamma\big), \qquad (11)$$

where a superscript * indicates that the treatment effect for the two models (with and without covariates) are different. Consequently, $\theta = \exp(\beta) \neq \exp(\beta^{\star}) \equiv \theta^{\star}$. If the true model has hazards given by equation (11), but we model the data with the proportional hazards model excluding covariates (i.e. using equation (10)), then the resulting estimator, say $\hat{\beta}$, does not consistently estimate $\beta^{\star}$.[42,43] Moreover, if equation (11) holds, the marginal model without covariates does not in general hold—a condition referred to as *non-collapsibility*. Note that in the normal linear model, the causal estimand is the same with or without covariates.

If the true model is given by equation (11), but we misspecify it as equation (10), we can still estimate a type of average hazard ratio based on maximizing the partial likelihood. Inferences are valid asymptotically if we use a robust variance estimator,[30] although the estimand can depend on the censoring distribution.[31–33] When equation (10) holds, the estimand θ is the population-level causal effect previously discussed, and when equation (11) holds the estimand $\theta^{\star}$ is an average of population-level estimands corresponding to different levels of X. If models with more and more baseline covariates added hold, then the estimand approaches an individual-level estimand (see the Supplementary Material Appendix A for an example definition of the individual-level estimand). Indeed, there is a continuum of estimands from the population-level estimand of equation (10) to the most subject-specific-like, which includes so many covariates that each individual has a unique X_i.[13]

Exploring treatment effects over time for survival data

For survival outcomes, interpretability is challenging when interest is in how treatment effects change or do not change over time. We explore this issue in two numerical examples.

Example 1: differential treatment effect

Consider a synthetic example based on Fay et al.[14] (Figure 2, Supplementary Material Appendix B), which was devised to highlight the difference between ϕ and ψ in equation (5). For this example, $Y_i = \{Y_i(0), Y_i(1)\} \sim F$, where F is a mixture of two bivariate normals, truncated to have all positive responses. Finally, we round the event times up to the nearest 10th, rendering the distributions discrete. We concoct a plausible story for the example. Part of the population (about 35%) have a severe form of the disease, and on them, the test treatment works very well (increasing the mean failure time from 1 to 6), but it additionally increases the variability. The other 65% have a mild form of the disease with mean survival time of 4 on control but slightly less on test. For them, the potential failure times are highly correlated. The bivariate probability mass function appears in Figure 1.

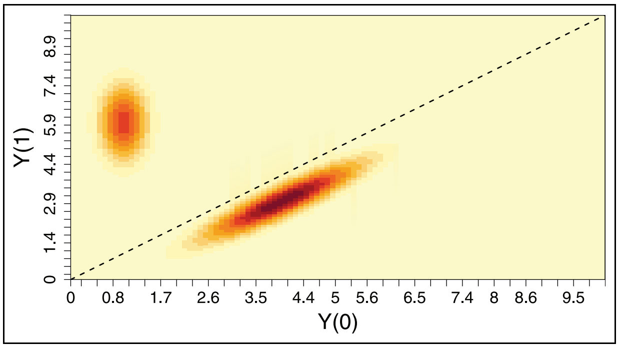

Figure 1. Example 1: potential outcome bivariate distribution, F, plotted in the form of its bivariate probability mass function, f. Darker values have more mass.

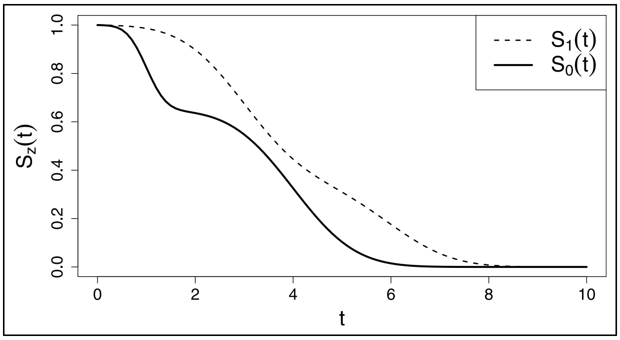

Figure 2. Example 1: potential outcome survival curves.

In a randomized trial with no baseline covariates to identify the latent classes, we can see only the marginal distributions. Figure 2 plots $S_1(t)$ and $S_0(t)$; Figure 3 plots $f_1(t)$ and $f_0(t)$; and Figure 4 plots the discrete hazard estimands $h_1(t)$ and $h_0(t)$, where $h_z(t_j) = f_z(t_j)/S_z(t_{j-1})$ for consecutive times t_{j-1} and t_j.

From Figure 2, it is evident that at the population-level the test treatment is effective at delaying the event. In Figure 3, the two subpopulations in each arm are apparent, and it also appears that the test treatment prolongs survival time at the population-level. In contrast, the hazard function plots in Figure 4 suggest that treatment is harmful at a range of values of t.

Now turn to the hazard ratio estimands. The discrete versions of $\theta(t)$, $\theta_{\mathrm{CHR}}(t)$, and the average hazard ratio estimands, are defined by replacing λ_z with h_z and replacing integrals by summations. The discrete version of the $\theta_{\mathrm{CoxAHR}}(t)$ is too difficult to calculate under non-proportional hazards, so we simulate it from the model with $n = 10,000$ and no other censoring except for all outcomes greater than t. The definition of the discrete version of the marginal causal hazard ratio is given in the Supplementary Material Appendix B.

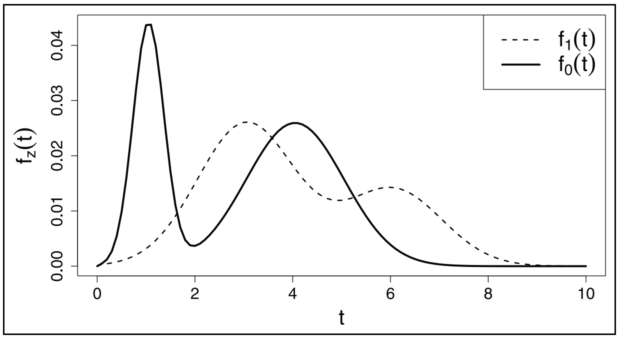

Figure 3. Example 1: potential outcome probability mass functions.

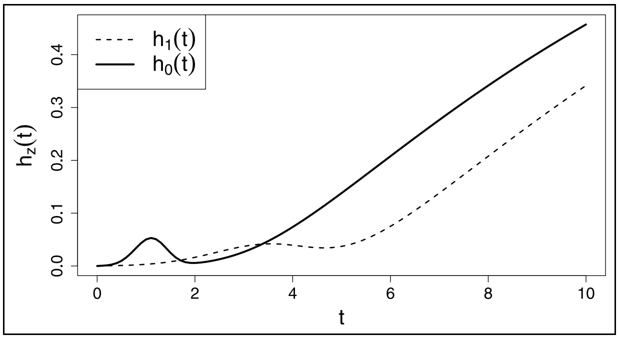

Figure 4. Example 1: potential outcome hazard curves.

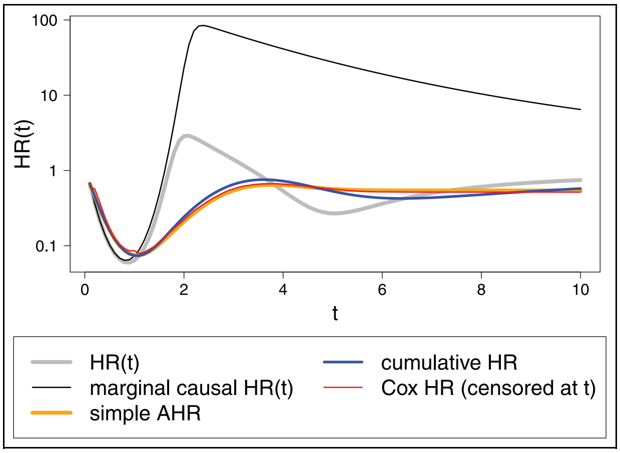

Figure 5. Example 1: hazard ratio estimands over time, except $\theta_{\text{CoxHR}}(t)$ is estimated from a simulated study with $n = 10,000$ with the end of study censoring changing such that $\tau = t$.

Figure 5 plots hazard ratio estimands over time. Notice first that there is a fundamental difference between the hazard ratio $\theta(t)$ and the marginal causal hazard ratio $\theta_{\text{mcHR}}(t)$ compared to the rest. The values $\theta(t)$ and $\theta_{\text{mcHR}}(t)$ represent two different types of instantaneous risk at each t, whereas the other three hazard ratios ($\theta_{\text{sAHR}}(t)$, θ_{CHR}, and $\theta_{\text{CoxAHR}}(t)$) measure a type of averaging of the hazard ratio up until time t. The version of $\theta_{\text{sAHR}}(t)$ that we use takes $w(t) = f_0(t)/f(t)$,[26] except done as a weighted average of $\theta(t)$, that is

$$\theta_{\text{sAHR}}(t_j) = \frac{\sum_{i=1}^{j} \left(\frac{h_1(t_i)}{h_0(t_i)}\right) f_0(t_i)}{\sum_{i=1}^{j} f_0(t_i)}.$$

In this example, the form of the average-type hazard ratio matters little, as the red, blue, and orange curves track closely, and we have previously shown that the cumulative hazard ratio, $\theta_{\text{CHR}}(t)$, has a population-level causal interpretation. The instantaneous risk estimands are very different, however. For example, at $t = 2$ we have $1 < \theta(2) < \theta_{\text{mcHR}}(2)$, whereas at $t = 4$ we have $\theta_{\text{mcHR}}(4) > 1 > \theta(4)$. To understand $\theta_{\text{mcHR}}(t)$ it is helpful to know F or $\Pr[\min(Y_i(0), Y_i(1)) > t]$, but neither is identifiable. Thus, knowing $\theta(t)$ does not even inform us about the direction of $\theta_{\text{mcHR}}(t)$.

Let us call the two classes of individuals latent class 1 (mean at $[1, 6]$; treatment helps) and latent class 2 (mean at $[4, 3]$; treatment harms). From Figure 1, we see that individuals who would survive past $t = 2$ regardless of treatment arm come almost entirely from latent class 2, where treatment harms. This is captured by $\theta_{\text{mcHR}}(t)$, which exceeds 1 for all $t > 2$, indicating treatment harm. For $\theta(t)$, we have different proportions in latent classes for the different conditioning sets. The proportion in latent class 1 (treatment helps) of those with $Y_i(0) \geqslant t_j$ is 0.35 at $t_j = 0$, and < 0.0005 for all $t_j > 2$. In contrast, the proportion in latent class 1 of those with $Y_i(1) \geqslant t_j$ is 0.35 at $t_j = 0$, 0.39 at $t_j = 2$, 0.77 at $t_j = 4$, and > 0.995 for all $t_j \geqslant 6$. Thus, the changing value of $\theta(t)$ is due to both the changing hazard rates within the latent classes and the changing proportion of the conditioning sets still at risk. This example illustrates the conceptual challenge in interpreting $\theta(t)$ as a causal estimand.

Example 2: vaccine trial

This example, based on a vaccine trial, demonstrates that in general, we cannot identify the difference between time-varying treatment effects on a homogeneous population and fixed treatment effects in a heterogeneous population.

We consider a model motivated by the first 6 months of data from the BNT162b2 mRNA COVID-19 vaccine trial described by Thomas et al.[44] Figure 2 of that article displays the cumulative incidence curves of the two arms: by 6 months, the placebo arm has a cumulative incidence of between 6% and 7%, whereas the vaccine arm has less than 1%. The vaccine efficacy is 91.2% from ⩾7 days after receipt of the second dose until about 6 months. Partitioning the study time after the receipt of the second dose into 3 periods, the vaccine efficacy estimates are 96.2% (⩾7 days to < 2 months), 90.1% (⩾2 months to < 4 months), and 83.7% (⩾4

Table 1. Hazard rates $(\times 10^6)$ for subgroups in the model that creates Figure 6.

Arm	Healthy	Frail	Ratios
Placebo	$\lambda_{\text{ph}} = 100$	$\lambda_{\text{pf}} = 12,000$	$\frac{\lambda_{\text{pf}}}{\lambda_{\text{ph}}} = 120$
Vaccine	$\lambda_{\text{vh}} = 4$	$\lambda_{\text{vf}} = 480$	$\frac{\lambda_{\text{vf}}}{\lambda_{\text{vh}}} = 120$
Ratios	$\frac{\lambda_{\text{vh}}}{\lambda_{\text{ph}}} = 0.04$	$\frac{\lambda_{\text{vf}}}{\lambda_{\text{pf}}} = 0.04$	

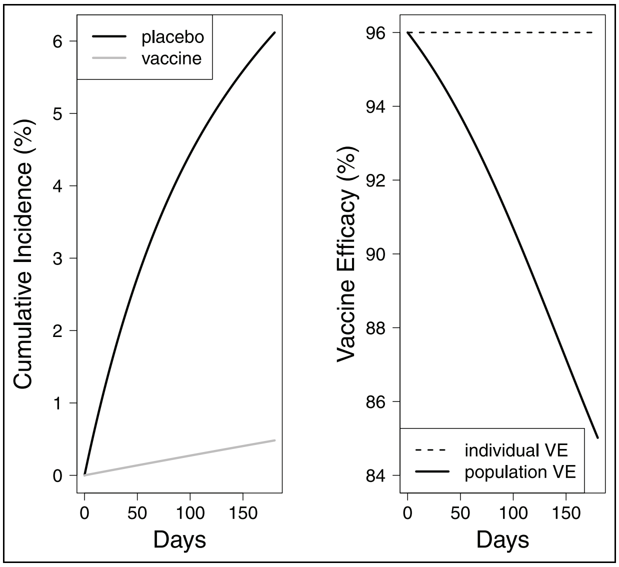

Figure 6. Example 2: vaccine example based on a mixture of four exponential curves, with hazard rates from Table 1.

months to about 6 months). Thus, the population vaccine efficacy is declining over time, but whether this is true of efficacy at the individual level is uncertain.

To examine these ideas, we propose two simple models. In Model 1, the population has two latent classes: 95% are healthy, and 5% are frail. We assume moreover that the vaccine efficacy is constant over time on every individual. We model the two potential outcomes for each class with an exponential model, with the hazard rates given in Table 1 (see the Supplementary Material Appendix C for more details). This model has a constant individual vaccine efficacy, $\left(1 - \frac{\text{hazard on vaccine}}{\text{hazard on placebo}}\right)$, of 96% for each individual regardless of health status. Furthermore, the effect of frailty is large and homogeneous within study arms; frail individuals have 120 times the instantaneous risk of healthy individuals, regardless of treatment. If we do not know the frailty status of the participants, we get the population cumulative incidence curves in the left panel of Figure 6 and the population vaccine efficacy in the right panel. These cumulative incidences and hazard ratios are similar to those in Figure 2 of Thomas et al.[44]

An important consideration is the distribution of frail individuals. By randomization, the fraction of frail individuals is 5% at time 0 in each arm. Assuming no censoring, at 50 days, the percentage is 2.8% in the placebo group and 4.9% in the vaccine arm; at 100 days, it is 1.6% in the placebo group and 4.8% in the vaccine arm; and at 150 days, it is 0.9% on placebo and 4.7% on vaccine. Thus, the composition of the risk sets is no longer balanced after randomization. If we know the frailty status of each individual and use that indicator in the Cox model, then we can consistently estimate the individual vaccine efficacy of 96% regardless of frailty status. This is the individual vaccine efficacy of Figure 6 that does not change over time.

In Model 2, we assume a homogeneous population in which the hazard ratio changes over time. In Supplementary Material Appendix C, we show that both models give the same expected cumulative incidence curve for each arm of the trial. Thus, if we only have time to event and treatment for each individual, we cannot identify from the data which of the two models is correct.

Although the models give identical cumulative incidence curves, they have very different causal interpretations. Under Model 1, the observed waning of efficacy is an artifact of the incompleteness of our information on individual health class, whereas under Model 2, the waning is real. Thus if Model 1 is true, it is more important to identify frail individuals and learn how to protect them. If Model 2 is true, it is more important to devise ways to boost the original vaccination over time as its efficacy wanes. Without information about frailty group membership, we cannot tell the difference between the two models. Importantly, in the COVID-19 vaccine example, we are confident that there is a large causal effect of the vaccine on the two arms, even if we cannot precisely identify the individual vaccine effects over time from an analysis of a large trial without using baseline covariates.

Discussion

We have reviewed causal estimands as comparators of potential outcomes in a randomized trial. We defined two types of causal estimands: With individual-level estimands, we first compare potential outcomes within an individual, then summarize those comparisons. With population-level estimands, we first summarize then compare the summaries. Difference-in-expectation estimands (e.g. Δ_{ATE} or $S_1(t^*) - S_0(t^*)$) are both individual-level and population-level estimands. Some other estimands—including causal rate ratios—have only a population-level interpretation.

Being an individual-level estimand is sufficient to have a causal interpretation, but it is not necessary, since population-level estimands that are not also individual-level estimands can have a causal interpretation as well, albeit a possibly less straightforward one. Importantly, the population-level hazard ratio does not have a straightforward causal interpretation. From one perspective, if we see a difference in the hazard functions, the randomization assures us that this difference did not arise because of confounding. The intervention has led to a change in the counterfactual distributions, and from this standpoint, the intervention has a causal effect. Furthermore, when the proportional hazards assumption holds for $t \in (0, t^{*}]$, since the hazard ratio is equivalent to the ratio of two log survival functions at t^{*}, and a comparison of those survival functions is a population-level causal estimand. From another perspective, however, the hazard ratio at t is not a causal estimand from a counterfactual standpoint, because the conditioning sets in the hazard function definition are not comparable for $t>0$ (unless there is no unaccounted for heterogeneity in the population, which is not a realistic assumption). Some would argue that to say "the intervention has a causal effect," but a parameter that describes that effect (such as the hazard ratio) "is not a causal estimand," is confusing. What is clear, however, is that $\theta(t)$ is *not* an individual-level causal estimand without strong homogeneity assumptions, so that we cannot in general interpret $\theta(t)$ as an average effect of changing the hazard of an event at t when on treatment compared to the hazard for that same individual on control. The reason for this is that the conditioning sets are still not equal between the two arms at each t such that $0<t<t^{*}$ and therefore one cannot assert, for example, that when $\theta(t) = \theta$, the treatment works equally well throughout the entire time course.

The causal hazard ratio at t of Martinussen et al.[24] (see also)[36] is the ratio of instantaneous risks at t given both potential outcomes are at least t. It is explicitly defined based on the same subpopulation in both arms and thus is more clearly a causal estimand. However, this is a theoretical estimand that is like a causal effect on the always-survivors (or a continuous-time survivor average causal effect)[36] and is not identifiable from trial data without additional assumptions.

Non-collapsibility of the Cox model complicates matters. When unadjusted and adjusted models are estimated, both cannot simultaneously be true. In a model where proportional hazards holds with covariates and treatment arm included, the estimated hazard ratio is a type of population-level causal estimand that is closer to (but not necessarily equal to) the individual-level causal estimand.

When the proportional hazards assumption does not hold, a useful estimand is the ratio of cumulative hazards (equivalently the ratio of log survival curves) comparing test to control at some specified time. This estimand has a population-level causal interpretation regardless of the proportional hazards assumption.[34]

Prentice and Aragaki[26] recommend that, when hazards are non-proportional, one use a statistic that is a type of average of the hazard ratios up until a designated time. Although this average hazard ratio is not a straightforward causal estimand, it is, at least in one example we explored, similar in magnitude to the ratio of cumulative hazards. It would be useful in future studies to explore whether this observation is generalizable to other data-generating processes.

Often in survival analyses, we seek to explore variation in treatment effects over time. We can get a population-level causal effect by plotting the ratio of cumulative hazards over time. This has less interpretational problems than $\theta(t)$, which represents the local hazard ratio at t and is likely to have different conditioning sets in the numerator and denominator. Nevertheless, a decreasing population cumulative hazard ratio does not imply that the individual hazard ratios are decreasing over time. This latter point is made clear by example 2, motivated by a real vaccine trial. There, we developed two data-generating models that result in the same pair of incidence curves for the two arms in the trial. These two models show that we cannot distinguish an effect on a homogeneous population where the individual-level hazard ratio is declining over time from an effect on a heterogeneous population with constant individual-level hazard ratios over time. That example suggests that to get closer to individual-level effects, we should include in our model baseline variables which may capture some of the frailty in the population. Thus, in that example, as with all the examples, one must bring to bear all relevant information in defining and interpreting trial-derived hazard ratios.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: F.L.'s research is supported by the National Institutes of Health (NIH) grant R01-HL168202. All statements in this article, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of NIH.

ORCID iDs

Michael P Fay https://orcid.org/0000-0002-8643-9625
Fan Li https://orcid.org/0000-0001-6183-1893

Supplemental material

Supplemental material for this article is available online.

References

1. Kahan BC, Cro S, Li F, et al. Eliminating ambiguous treatment effects using estimands. *Am J Epidemiol* 2023; 192: 987–994.
2. Imbens GW and Rubin DB. *Causal inference in statistics, social, and biomedical sciences*. Cambridge: Cambridge University Press, 2015.
3. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B* 1972; 34: 187–202.
4. Hernán MA. The hazards of hazard ratios. *Epidemiology* 2010; 21: 13–15.
5. Aalen OO, Cook RJ and Røysland K. Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal* 2015; 21(4): 579–593.
6. Stensrud MJ, Aalen JM, Aalen OO, et al. Limitations of hazard ratios in clinical trials. *Eur Heart J* 2019; 40: 1378–1383.
7. Martinussen G. Causality and the Cox regression model. *Ann Rev Stat Appl* 2022; 9: 49–59.
8. Hernán M and Robins J. *Causal inference: what If*. Boca Raton, FL: Chapman & Hall/CRC, 2020.
9. Frangakis CE and Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; 58: 21–29.
10. Angrist JD, Imbens GW and Rubin DB. Identification of causal effects using instrumental variables. *J Am Stat Assoc* 1996; 91: 444–455.
11. Zhang JL and Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by "death."*J Educ Behav Stat* 2003; 28(4): 353–368.
12. Zeger SL, Liang KY and Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 1988; 44(4): 1049–1060.
13. Hauck WW, Anderson S and Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials. *Control Clin Trials* 1998; 19(3): 249–256.
14. Fay MP, Brittain EH, Shih JH, et al. Causal estimands and confidence intervals associated with Wilcoxon-Mann-Whitney tests in randomized experiments. *Stat Med* 2018; 37: 2923–2937.
15. Mao L. On causal estimation using U-statistics. *Biometrika* 2018; 105(1): 215–220.
16. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; 53(282): 457–481.
17. Klein JP, Logan B, Harhoff M, et al. Analyzing survival curves at a fixed point in time. *Stat Med* 2007; 26: 4505–4519.
18. Fay MP, Proschan MA and Brittain E. Combining one-sample confidence procedures for inference in the two-sample case. *Biometrics* 2015; 71(1): 146–156.
19. Robins JM and Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS clinical trial with Inverse Probability of Censoring Weighted (IPCW) log-rank tests. *Biometrics* 2000; 56(3): 779–788.
20. Cheng C, Li F, Thomas LE, et al. Addressing extreme propensity scores in estimating counterfactual survival functions via the overlap weights. *Am J Epidemiol* 2022; 191(6): 1140–1151.
21. Karrison TG. Use of Irwin's restricted mean as an index for comparing survival in different treatment groups — interpretation and power considerations. *Control Clin Trials* 1997; 18(2): 151–167.
22. Zhao L, Tian L, Uno H, et al. Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clin Trials* 2012; 9(5): 570–577.
23. Wei LJ. The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Stat Med* 1992; 11(14–15): 1871–1879.
24. Martinussen T, Vansteelandt S and Andersen PK. Subtleties in the interpretation of hazard contrasts. *Lifetime Data Anal* 2020; 26(4): 833–855.
25. Schemper M, Wakounig S and Heinze G. The estimation of average hazard ratios by weighted Cox regression. *Stat Med* 2009; 28(19): 2473–2489.
26. Prentice RL and Aragaki AK. Intent-to-treat comparisons in randomized trials. *Stat Sci* 2022; 37: 380–393.
27. Xu R and O'Quigley J. Estimating average regression effect under non-proportional hazards. *Biostatistics* 2000; 1(4): 423–439.
28. Kalbfleisch JD and Prentice RL. Estimation of the average hazard ratio. *Biometrika* 1981; 68(1): 105–112.
29. Oakes D. On the win-ratio statistic in clinical trials with multiple types of event. *Biometrika* 2016; 103(3): 742–745.
30. Lin DY and Wei LJ. The robust inference for the Cox proportional hazards model. *J Am Stat Assoc* 1989; 84: 1074–1078.
31. Nguyen VQ and Gillen DL. Robust inference in discrete hazard models for randomized clinical trials. *Lifetime Data Anal* 2012; 18(4): 446–469.
32. Struthers CA and Kalbfleisch JD. Misspecified proportional hazard models. *Biometrika* 1986; 73(2): 363–369.
33. Rufibach K. Treatment effect quantification for timetoevent endpoints—estimands, analysis strategies, and beyond. *Pharm Stat* 2019; 18(2): 145–165.
34. Wei G and Schaubel DE. Estimating cumulative treatment effects in the presence of nonproportional hazards. *Biometrics* 2008; 64(3): 724–732.
35. Vansteelandt S, Dukes O, Van Lancker K, et al. Assumption-lean Cox regression. *J Am Stat Assoc* 2022; 119: 475–484.
36. Axelrod R and Nevo D. A sensitivity analysis approach for the causal hazard ratio in randomized and observational studies. *Biometrics* 2023; 79: 2743–2756.
37. Tsiatis AA, Davidian M, Zhang M, et al. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Stat Med* 2008; 27: 4658–4677.
38. Wang B, Ogburn EL and Rosenblum M. Analysis of covariance in randomized trials: more precision and valid confidence intervals, without model assumptions. *Biometrics* 2019; 75(4): 1391–1400.
39. Zeng S, Li F, Wang R, et al. Propensity score weighting for covariate adjustment in randomized clinical trials. *Stat Med* 2021; 40(4): 842–858.

40. Karrison T and Kocherginsky M. Restricted mean survival time: does covariate adjustment improve precision in randomized clinical trials. *Clin Trials* 2018; 15(2): 178–188.
41. Mehrotra DV and Marceau West R. Survival analysis using a 5-step stratified testing and amalgamation routine (5-STAR) in randomized clinical trials. *Stat Med* 2020; 39(30): 4724–4744.
42. Gail MH, Wieand S and Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika* 1984; 71: 431–444.
43. Neuhaus JM and Jewell NP. A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika* 1993; 80: 807–815.
44. Thomas SJ, Moreira Jr ED, Kitchin N, et al. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine through 6 months. *N Engl J Med* 2021; 385(19): 1761–1773.