*Statistical Science*
2022, Vol. 37, No. 3, 380–393
https://doi.org/10.1214/21-STS830


# Intention-to-Treat Comparisons in Randomized Trials

**Ross L. Prentice and Aaron K. Aragaki**

*Abstract.* Intention-to-treat (ITT) comparisons have a central place in reporting on randomized controlled trials, though there are typically additional analyses of interest such as those making adjustments for nonadherence. In our ITT reporting of results from the Women's Health Initiative (WHI) randomized trials, we have relied primarily on highly flexible hazard ratio (Cox) regression methods. However, these methods, especially the proportional hazards special case, have been criticized for being difficult to interpret and frequently oversimplified, and for not being consistent with modern causality theories using potential outcomes. Here we address these topics and extend our use of hazard rate methods for ITT comparisons in the WHI trials.

*Key words and phrases:* Causality, Cox model, failure time data, regression, restricted mean survival time.

## 1. INTRODUCTION

Prior to 1972 censored time-to-response data, often referred to as 'survival' data, were typically analyzed using fully parametric models, such as exponential or Weibull models (e.g., Feigl and Zelen, 1965) or were analyzed using a Mantel–Haenszel (1959) stratified odds ratio approach. The seminal paper by Cox (1972) quickly changed this landscape. Cox's hazard ratio regression model extended the parametric models by including a nonparametric baseline hazard rate factor, along with an exponential form parametric hazard ratio factor. In doing so, it also extended the flexibility of the Mantel–Haenszel estimator to a full regression model for the instantaneous odds ratio, or hazard ratio (HR). The regression parameter in the Cox model was shown subsequently to be estimated in a semiparametric efficient manner by Cox's (1975) partial likelihood estimator, at least if regression variates are time-independent (Begun et al., 1983). Some decades of research followed that augmented the class of Cox models to include stochastic time-varying regression variables and stratified time-varying baseline hazard rates, to allow multiple failure types and multivariate failure time outcomes, and to accommodate missing and mismeasured covariate data, among many other extensions. Novel research efforts focused also on asymptotic properties of estimators of parameters in hazard ratio regression models, including asymptotic properties under a variety of study subject selection and censoring patterns, as well as under a variety of sampling procedures for (expensive) covariate history assembly. Many of these developments have been summarized in books along the way (e.g., Andersen et al., 1993, Kalbfleisch and Prentice, 2002, O'Quigley, 2008, Aalen, Borgan and Gjessing, 2010, Cook and Lawless, 2018, Prentice and Zhao, 2019).

As the Cox model, especially its proportional hazards special case, became a prominent data analytic tool for follow-up studies, some issues emerged. Of course it was recognized early that some bias may arise for HR estimators under departure from the assumed parametric HR model. For example, Struthers and Kalbfleisch (1986) investigated biases associated with fitting a proportional hazards model when an accelerated failure time model obtains. See also Lagakos and Schoenfeld (1984) for studies of the impact of nonproportionality and missing covariates in a randomized trial setting. Hernán (2010) noted that an estimated (constant) HR in a two-sample comparison is difficult to interpret if the HR changes over follow-up time, and the characterized HR estimates at a specific follow-up time as having 'built-in selection bias' due to differential failure rates among susceptibles at times prior to $t$. Aalen, Cook and Røysland (2015) noted further that the HR is 'not a quantity that admits a causal interpretation in the case of unmodeled heterogeneity' even if the HR is correctly modeled. Aalen et al. summarize by suggesting that 'modeling frameworks more compatible with

*Ross L. Prentice is PhD, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA and Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA (e-mail: rprentic@whi.org). Aaron K. Aragaki is MS, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, USA.*

causal reasoning may be preferable for estimation in general.'

Over much of this same time period a major development arose to formalize statistical aspects of causality in terms of potential, or counterfactual, outcomes; that is, outcomes that would arise if a study subject could be assigned to each of two treatment groups in a randomized controlled trial, with early influential contributions by Rubin (1974, 1978) and Robins (1986). Much of this work was motivated by a desire to strengthen observational study analyses through emulation of randomized trials, by measuring additional variables which, when conditioned upon, may be able to yield a state of exchangeability between treatments or exposure groups to be compared. The resulting methods were proposed also to correct for noncompliance in randomized trials (e.g., Robins, 1994), and ultimately as a framework for judging ITT analyses of randomized trials according to whether or not the estimands employed can be represented as averages of counterfactual differences (e.g., Aalen, Cook and Røysland, 2015). For example, the preference for such methods appears to have contributed to recent interest in the use of restricted mean survival time (RMST) contrasts rather than hazard ratio analyses (e.g., Uno et al., 2015, Zhao et al., 2016), and to accelerated failure time models rather than Cox regression models (e.g., Hernán and Robins, 2020, Chapter 17) for clinical trial reporting with failure time outcomes.

Here we present a defense of hazard rate modeling, and Cox regression in particular, for the ITT reporting of causal effects in randomized controlled trials. We do so in the context of the massive Women's Health Initiative (WHI) randomized, placebo controlled hormone therapy trials in which we have been engaged for many years. The recent book by Hernán and Robins (2020) provides a very readable version of much of the literature on causality using a potential outcomes framework, and we will draw on this work as an authoritative source in some of our arguments.

## 2. WHI MENOPAUSAL HORMONE THERAPY TRIALS

The WHI hormone therapy trials of conjugated equine estrogens 0.625 mg/d continuous and medroxyprogesterone acetate 2.5mg/d continuous (CEE+MPA) among 16,608 postmenopausal US women with uterus, and the companion trial of the same estrogens preparation (CEE) among 10,739 women who were post-hysterectomy, led to a sea change in clinical practice, with about 70% of the approximately 6 million women using this CEE+MPA regimen, and about 40% of the approximately 8 million women using this CEE regimen, stopping usage abruptly when the CEE+MPA trial was stopped early and results were released following an average 5.6 year intervention period (Rossouw et al., 2002). The trigger for early stoppage was a significant elevation in breast cancer risk, in conjunction with some elevations also in cardiovascular disease incidence, for which major reductions had been hypothesized based on a substantial observational literature. The CEE trial was also stopped early in 2004 following an average 7.2 year intervention period (Anderson et al., 2004), largely based on a stroke elevation of similar magnitude to that for CEE+MPA. Reductions in postmenopausal breast cancer incidence in the US (e.g., Ravdin et al., 2007) and elsewhere were reported soon thereafter and reductions in US health care costs from the reduction in use of CEE+MPA have been estimated at $37.1 billion through 2012 (Roth et al., 2014).

This brief account illustrates the weight that is attached to well-conducted randomized trials by clinicians and regulators, as derives from the independence between intervention group assignment and baseline risk factors, whether recognized as such or not, in randomized trials. Assuming equal outcome ascertainment between randomized groups, and the absence of post-randomization confounding by unplanned changes in the study population during follow-up, it is logical to regard the outcome patterns that emerge as being caused by the treatments under study. As emphasized by Cox (1992) these types of causal arguments may be able to be strengthened through elucidation of biological mechanisms that attend differences in outcome patterns between randomization groups.

The Cox regression method, with its time-dependent covariate and stratification features has been the 'workhorse' for reporting on the WHI hormone therapy trials over many years, as nonintervention follow-up still continues, with median follow-up now in excess of 20 years. An early analysis of the CHD primary efficacy outcomes (Manson et al., 2003) listed HR estimates (95% CIs) for CEE+MPA of 1.81 (1.09, 3.01), 1.34 (0.82, 2.18), 1.27 (0.64, 2.50) 1.25 (0.74, 2.12), 1.45 (0.81, 2.59) and 0.70 (0.42, 1.14) for years 1, 2, 3, 4, 5 and $\geq$6 years from randomization. The early elevation in CHD rates had not been recognized in the preceding observational study reports. When the CEE+MPA trial data were analyzed jointly with a corresponding sub-cohort from a WHI Observations Study ($n = 93{,}676$), a good agreement in HR pattern between cohorts was found after allowing for time-dependence in the HR function (Prentice et al., 2005). CHD contrasts in the CEE trial were similar, but somewhat less pronounced (Hsia et al., 2006), and again agreed well with HRs from the pertinent component of the OS after allowing for HR function time-dependence (Prentice et al., 2006).

An elevation in (invasive) breast cancer risk with postmenopausal hormones was anticipated from observational studies, leading to the specification of breast cancer as primary safety outcome in both trials. An observed breast cancer risk elevation was a key element of the early stoppage decision in 2002 for the CEE+MPA trial (Rossouw

et al., 2002). Following a short period of lower breast cancer incidence in the active CEE+MPA group, perhaps due to influences of the intervention on mammographic density with resulting delay in breast cancer diagnosis, the breast cancer HR increased quite substantially with time from randomization (e.g., Chlebowski et al., 2009) and continued to be elevated with long-term follow-up (e.g., Manson et al., 2013). In contrast, and unexpectedly, the breast cancer HR was reduced in the CEE trial (Anderson et al., 2004), and continued to be reduced over long-term follow-up (e.g., Chlebowski et al., 2020). The major difference in breast cancer results from the two trials is a continuing source of study and debate. The issues under consideration primarily relate to variation in hormone therapy influences according to such participant characteristics as age at starting menopausal hormones (Prentice et al., 2020), gap time from menopause to start of hormone therapy (Prentice et al., 2009), and whether or not hysterectomy was accompanied by bilateral oopherectomy (Manson et al., 2019). We believe that these timing issues are an appropriate focus for understanding the causality and magnitude of effects of menopausal hormonal therapy on CHD and breast cancer, and that our (sometimes) oversimplified HR modeling using the Cox model has provided reliable insights into these effects.

## 3. HAZARD RATE CONTRASTS IN RANDOMIZED CONTROLLED TRIALS

### 3.1 Estimands

A general regression notation will be used for censored time-to-response, or failure time, data. Consider a univariate failure time variate $T > 0$, subject to right censoring by variate $C \geq 0$, along with a covariate process $Z$ that may be evolving over time in a study population. Denote by $z(s) = \{z_1(s), z_2(s), \ldots\}$ the covariate value at time $s \geq 0$, and by $Z(t) = z(0) \vee \{z(s), 0 < s < t\}$ the covariate history prior to time $t$. Hazard rates are fundamental to the representation and analysis of censored failure time data, since it is precisely these rates that are identifiable under an independent censoring assumption. The hazard rate, given $Z$, at follow-up time $t$ can be written

$$\Lambda\{dt; Z(t)\} = P\{t \leq T < t + dt; T \geq t, Z(t)\}, \tag{1}$$

and independent censorship requires that $C \geq t$ can be added to the conditioning event without changing the hazard rate for any $t \geq 0$ and $Z(t)$. The corresponding (cumulative) hazard function, using Stieltjes integration, is given by

$$\Lambda\{t; Z(t)\} = \int_0^t \Lambda\{ds; Z(s)\}. \tag{2}$$

A randomized controlled trial typically has $Z(t) \equiv z$ for all $t \geq 0$, where $z = 0$ in a control group, and $z = 1$ in an active treatment group. ITT comparisons between randomized groups can be based on test statistics, such as the logrank test, or on comparisons of various functions of hazard rates. For example, ITT comparisons having causal interpretations can derive from differences between groups of survival functions

$$F(t; z) = \prod_0^t \{1 - \Lambda(ds; z)\}, \tag{3}$$

where $\prod$ denotes product integral. A causal interpretation also derives from the randomized trial design for many other functional estimands, based on hazard functions. Such a causal interpretation has been noted also from a potential outcome perspective. For example, (Hernán and Robins, 2020, p. 7) write for a response variable not subject to censoring that '... a population causal effect may be defined as a contrast of functionals, including medians, variances, hazards, or cdfs of counterfactual outcomes'. Similarly, Martinussen, Vansteelandt and Andersen, 2020, in a censored failure time setting, note that various functional contrasts derived from hazard rates, have a causal interpretation using counterfactuals.

One such functional contrast that has been advocated for use in a randomized controlled trials is the restricted mean survival time difference

$$\text{RMST}(t; z = 1) - \text{RMST}(t; z = 0), \quad \text{for } t \geq 0, \tag{4}$$

where

$$\text{RMST}(t; z) = \int_0^t F(s; z) = \int_0^t \prod_0^s \{1 - \Lambda(ds; z)\},$$

which has an attractive interpretation as the difference between groups in the expected time without failure in $[0, t]$ (e.g., Uno et al., 2015, Zhao et al., 2016). However, in disease prevention trials, such as the WHI hormone therapy trials, the restricted mean difference may be quite small during follow-up, perhaps reducing its value as a results communication tool. Also (4) may depend more strongly on eligibility and exclusionary criteria in a trial setting, than does, for example, a functional based on relative hazard rates between treatment groups.

There is also a literature on average hazard ratio (AHR) estimands (e.g., Kalbfleisch and Prentice, 1981, Schemper, Wakounig and Heinze, 2009). For example, when there is a specific control group one can define an AHR contrast by

$$\text{AHR}(t) = \int_0^t \frac{\Lambda(ds; z = 1)}{\Lambda(ds; z = 0)} P(ds; z = 0) \quad \text{for } t > t_0, \tag{5}$$

where $P$ denotes the failure time probability distribution in the control group, conditioned on $[0, t]$, and $t_0 > 0$ is chosen to ensure a positive control group failure probability in $[0, t]$.

### 3.2 Estimation

The estimands described above can each be estimated nonparametrically by inserting Nelson–Aalen estimators of hazard rates (2) for $t \geq 0$. Specifically, based on an independent random sample, $\{S_i = T_i \wedge C_i, \delta_i = I[T_i = S_i], z_i\}$, $i = 1, \ldots, n$, where $I(\cdot)$ denotes an indicator function, this hazard function estimator $\{\hat{\Lambda}(t; z); 0 \leq t \leq \tau\}$, for $\tau$ in the support of the observed times $T_i \wedge C_i, i = 1, \ldots, n$, can be written as

$$\hat{\Lambda}(t; z) = \int_0^t \sum_{i=1}^n I(z = z_i) N_i(ds) \Big/ \sum_{i=1}^n I(z = z_i) Y_i(s), \tag{6}$$

where $N_i(ds) = 1$ if $S_i = s$ and $\delta_i = 1$, and $N_i(ds) = 0$ otherwise, and $Y_i(s) = 1$ if $s \leq S_i$ and $Y_i(s) = 0$ otherwise. This estimator has well established asymptotic convergence properties, and these lead to standard $n^{-1/2}$ convergence rates for many related compact differentiable transformations, including Kaplan–Meier survival function estimators, given by

$$\hat{F}(t; z) = \prod_0^t \{1 - \hat{\Lambda}(ds; z)\},$$

as well as estimators given by

$$\hat{\text{AHR}}(t) = \{1 - \hat{F}(t; z = 0)\}^{-1} \times \int_0^t \hat{F}(s; z = 0) \hat{\Lambda}(ds; z = 1), \quad \text{for } t > t_0$$

for the average hazard ratio estimand (5). The asymptotic results alluded to above imply bootstrap applicability for each of these functionals, as well as for the RMST difference function.

### 3.3 Illustration

Table 1 shows the number of study participants developing clinical outcomes during the intervention phase, and during the subsequent nonintervention follow-up phase of the WHI hormone therapy trials, for CHD, breast cancer, and for several other important clinical outcomes that together constituted a 'global index' used for trial monitoring and reporting.

Figure 1 shows AHR estimators and RMST contrasts for the CEE+MPA trial for CHD (upper) and breast cancer (lower), along with pointwise 95% confidence intervals based on 1000 bootstrap samples. Both displays start at 1 year post-randomization and continue through year 16. The AHR estimate for CHD shows an early elevation but ceases to be significantly elevated after about 6 years from randomization, whereas the RMST contrast doesn't clearly show a randomization influence. For breast cancer the AHR function can be observed to be elevated by about 30% over the first six years from randomization, and evidence for an elevation becomes stronger with longer, mostly nonintervention, follow-up. A reduction in RMST for breast cancer in the active treatment group is not evident until about 12 years following randomization.

Supplementary Figure 1 (Prentice and Aragaki, 2022) shows corresponding estimates for the CEE trial. Neither estimator identifies a clear influence of randomization to active CEE on CHD risk, while both estimate a breast cancer risk reduction, starting at about 5 years from randomization for the AHR estimator, and at about 8 years for the RMST contrast. From these analyses, one might speculate that AHR functions may be particularly useful for identifying early differences between randomization groups, especially with rare outcomes.

The failure times in these analyses were times from randomization to the occurrence of the outcome under analysis, while the potential censoring times were the earliest of time to the end of the follow-up period, time to earlier loss to follow-up (which occurred rarely), and time to death. Some authors (e.g., Fine, Jiang and Chappell, 2001) regard time to death as a competing risk and entertain times to CHD, or to breast cancer, that could have occurred had the participant not died from some other cause. This approach leads to identifiability issues and strong additional assumptions for estimation of targeted quantities. Instead, in our WHI analyses we regard the hazard rates (1) as implicitly conditioning on the continued survival of the study subject to time $t$. A death from a competing cause then simply shortens the time period for the individual's contribution to hazard rate estimation, without requiring additional assumption. This approach, of course, has implications for the interpretation of treatment group contrasts, but has the advantage of retaining randomization-based causal interpretations for functionals based on comparisons among identifiable hazard functions.

As noted by Hernán (2010), hazard rates are somewhat complex to describe, and hazard rate comparisons at specific times $t > 0$ typically do not involve comparisons between randomized groups. However, the hazard rate contrasts at a particular follow-up time, and contrasts for the entire hazard function, have values that can be attributed to the randomized group assignment. Though there is room for discussion, we think from the AHR component of Figure 1 that a summary such as 'persons like you who are randomly assigned to CEE+MPA have about an 80% increase in risk for CHD during the first year compared to similar persons assigned to placebo, but an overall elevation in CHD is not evident over a longer-term intervention and follow-up period' provides a useful and appropriate communication for clinical application. Similarly, for CEE+MPA and breast cancer a clinician could appropriately summarize the AHR display of Figure 1 as 'there may be a small early reduction in breast cancer diagnosis among women assigned to active CEE+MPA versus

TABLE 1

*Hazard ratios and 95% confidence intervals under a simple proportional hazards assumption, and corresponding inverse variance weighted estimators over the phases of the Women's Health Initiative menopausal hormone therapy trials, for major clinical outcomes*

| | Intervention Phase | | | | Post-Intervention | | | | | Proportional Hazards[1] | | | Inverse Variance Weighted[2] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Clinical outcome | Cases active | Cases placebo | $\hat{\beta}$ | $SE\hat{\beta}$ | Cases active | Cases placebo | $\hat{\beta}$ | $SE\hat{\beta}$ | Test of Equality $p$-value[3] | HR | 95% CI | | HR | 95% CI | |
| **CEE+MPA Trial** | | | | | | | | | | | | | | | |
| Primary Outcomes | | | | | | | | | | | | | | | |
| Coronary heart disease | 196 | 159 | 0.162 | 0.107 | 514 | 493 | 0.015 | 0.063 | 0.23 | 1.05 | 0.95 | 1.17 | 1.05 | 0.95 | 1.17 |
| Invasive breast cancer | 205 | 155 | 0.211 | 0.107 | 369 | 277 | 0.258 | 0.080 | 0.72 | 1.27 | 1.12 | 1.44 | 1.27 | 1.12 | 1.44 |
| Secondary outcomes | | | | | | | | | | | | | | | |
| Stroke | 159 | 110 | 0.308 | 0.125 | 420 | 382 | 0.056 | 0.071 | 0.08 | 1.13 | 1.00 | 1.27 | 1.12 | 1.00 | 1.27 |
| Pulmonary embolism | 87 | 41 | 0.683 | 0.190 | 148 | 158 | −0.118 | 0.115 | <0.001 | 1.11 | 0.92 | 1.35 | 1.10 | 0.91 | 1.33 |
| Colorectal cancer | 50 | 76 | −0.493 | 0.182 | 128 | 121 | 0.010 | 0.127 | 0.02 | 0.85 | 0.70 | 1.05 | 0.86 | 0.70 | 1.05 |
| Endometrial cancer | 27 | 30 | −0.182 | 0.266 | 70 | 97 | −0.382 | 0.157 | 0.52 | 0.72 | 0.55 | 0.94 | 0.72 | 0.55 | 0.94 |
| Hip fracture | 53 | 75 | −0.402 | 0.180 | 341 | 346 | −0.051 | 0.077 | 0.07 | 0.90 | 0.78 | 1.03 | 0.90 | 0.78 | 1.03 |
| All-cause mortality | 250 | 238 | −0.027 | 0.091 | 1620 | 1555 | 0.003 | 0.036 | 0.76 | 1.00 | 0.94 | 1.07 | 1.00 | 0.94 | 1.07 |
| **CEE Trial** | | | | | | | | | | | | | | | |
| Primary Outcomes | | | | | | | | | | | | | | | |
| Coronary heart disease | 205 | 222 | −0.055 | 0.097 | 316 | 328 | −0.011 | 0.079 | 0.73 | 0.97 | 0.86 | 1.10 | 0.97 | 0.86 | 1.10 |
| Invasive breast cancer | 103 | 135 | −0.244 | 0.131 | 128 | 156 | −0.194 | 0.119 | 0.77 | 0.81 | 0.68 | 0.96 | 0.81 | 0.68 | 0.96 |
| Secondary Outcomes | | | | | | | | | | | | | | | |
| Stroke | 174 | 130 | 0.330 | 0.116 | 225 | 262 | −0.114 | 0.091 | 0.003 | 1.06 | 0.92 | 1.22 | 1.06 | 0.92 | 1.22 |
| Pulmonary embolism | 52 | 39 | 0.303 | 0.212 | 101 | 111 | −0.064 | 0.138 | 0.15 | 1.05 | 0.84 | 1.31 | 1.05 | 0.83 | 1.31 |
| Colorectal cancer | 65 | 58 | 0.139 | 0.181 | 54 | 60 | −0.102 | 0.188 | 0.35 | 1.02 | 0.79 | 1.32 | 1.02 | 0.79 | 1.32 |
| Hip fracture | 48 | 74 | −0.407 | 0.185 | 160 | 155 | 0.044 | 0.113 | 0.04 | 0.92 | 0.76 | 1.11 | 0.92 | 0.77 | 1.12 |
| All-cause mortality | 301 | 299 | 0.034 | 0.082 | 957 | 978 | −0.006 | 0.046 | 0.67 | 1.00 | 0.93 | 1.08 | 1.00 | 0.93 | 1.08 |

[1]Possibly oversimplified proportional hazards (Cox) model over cumulative follow-up, with baseline hazard function stratified on age, dietary modification trial randomization status, prior history of disease under analysis (if applicable), and study phase (time-dependent).

[2]Same as proportional hazards model described in footnote 1, except treatment HRs are allowed to differ in intervention and post-intervention phases, and summary HR (95% CI) is from average inverse variance weighted linear combination of log HRs from the two phases.

[3]Stratified logrank test significance level ($p$-value) for testing equality of HRs between the two study phases.

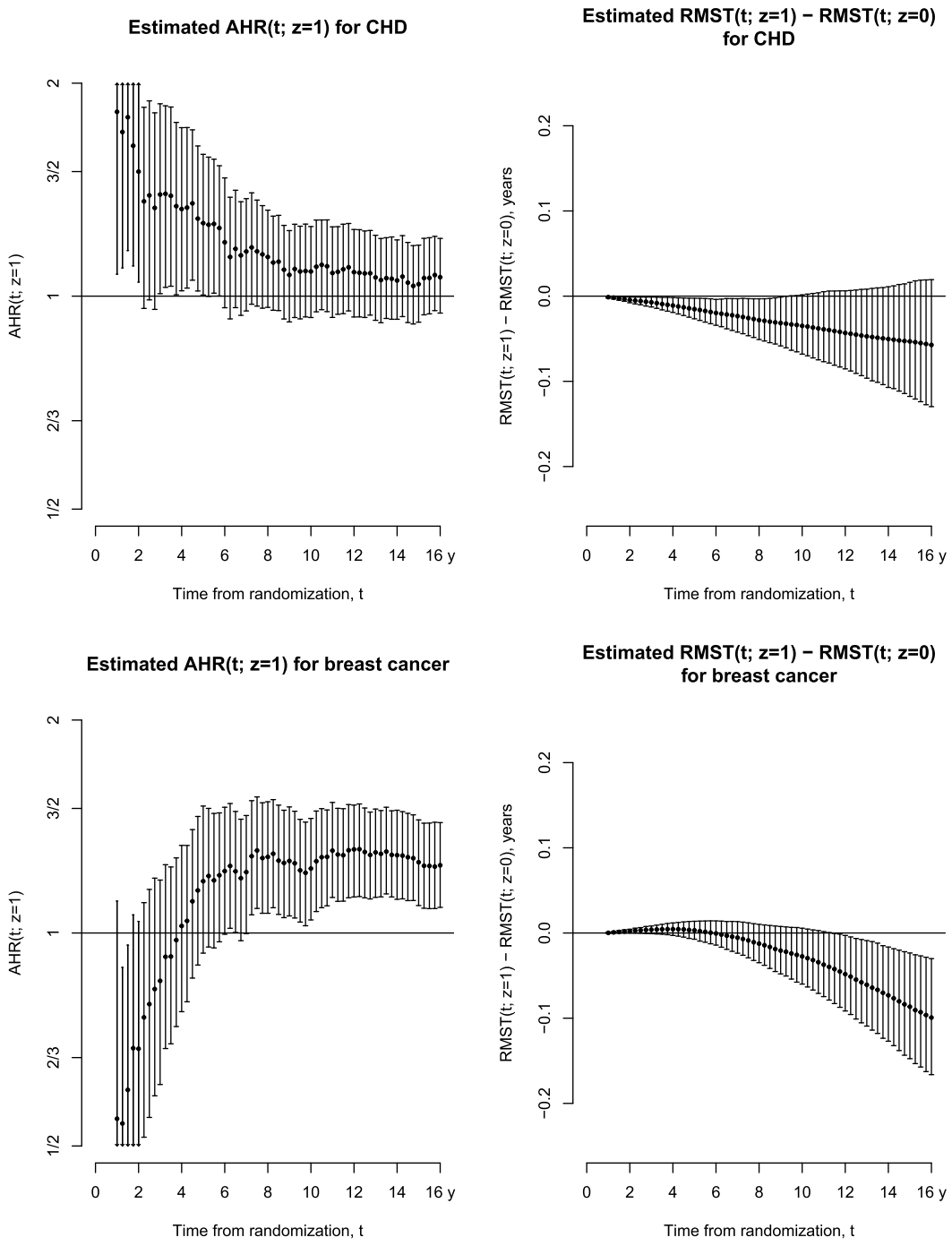


FIG. 1. *Nonparametrically estimated average hazard ratios* (*AHRs*) *and restricted mean survival time differences* (*RMST active–RMST placebo*) *with nonparametric bootstrap* 95% *confidence intervals* (1000 *bootstrap samples*) *for coronary heart disease* (*CHD*) *and breast cancer in the Women's Health Initiative estrogen plus progestin* (*CEE+MPA*) *trial among* 16,608 *postmenopausal US women with uterus.*

placebo, possibly due to diagnosis delays in the intervention group, but by 6 years the breast cancer risk is elevated by about 30% on average in the intervention group, and remains elevated over longer-term intervention and follow-up'.

There are, of course, many other possibilities beyond AHR and RMST contrasts for making useful ITT comparisons between randomized groups in clinical trials with failure time outcomes. In particular there has been much study and comparison null hypothesis tests for this purpose. The left side of Table 2 shows $p$-values for an RMST contrast at a 'landmark time' defined as the smaller of the maximum intervention phase follow-up times in the two groups, and for a traditional intervention phase logrank test which is a score test of null hypothesis against a proportional hazards alternative under which $\Lambda(dt; z = 1)/\Lambda(dt; z = 0)$ is constant as a function of $t$. The $p$-values for the two tests are similar for

TABLE 2

*Significance levels (p-values) for null hypotheses test over the intervention phase of the Women's Health Initiative hormone therapy trials*

| | | | HR model with $x(t) = (z, zt)$ | | | | | |
|---|---|---|---|---|---|---|---|---|
| Clinical Outcome | RMST $P$value[1] | $P$ value | Logrank $\hat{\beta}_1$ | SE | $\hat{\beta}_2$ | SE | PH Test $P$ value[2] | 2DF $P$ value[3] |
| | | | | CEE+MPA Trial | | | | |
| Primary Outcomes | | | | | | | | |
| Coronary heart disease | 0.24 | 0.19 | 0.570 | 0.203 | −0.136 | 0.054 | 0.01 | 0.02 |
| Invasive breast cancer | 0.17 | 0.05 | −0.340 | 0.232 | 0.163 | 0.062 | 0.008 | 0.004 |
| Secondary Outcomes | | | | | | | | |
| Stroke | 0.01 | 0.02 | 0.254 | 0.253 | 0.013 | 0.066 | 0.85 | 0.06 |
| Pulmonary embolism | <0.001 | <0.001 | 1.272 | 0.372 | −0.181 | 0.094 | 0.05 | <0.001 |
| Colorectal cancer | 0.009 | 0.007 | −0.361 | 0.359 | −0.041 | 0.100 | 0.69 | 0.02 |
| Endometrial cancer | 0.51 | 0.49 | 0.246 | 0.579 | −0.124 | 0.150 | 0.40 | 0.56 |
| Hip fracture | 0.03 | 0.02 | −0.387 | 0.369 | −0.010 | 0.100 | 0.92 | 0.06 |
| All-cause mortality | 0.74 | 0.75 | 0.110 | 0.205 | −0.038 | 0.051 | 0.45 | 0.71 |
| | | | | CEE Trial | | | | |
| Primary Outcomes | | | | | | | | |
| Coronary heart disease | 0.74 | 0.59 | 0.178 | 0.199 | −0.055 | 0.042 | 0.19 | 0.36 |
| Invasive breast cancer | 0.16 | 0.06 | −0.505 | 0.283 | 0.063 | 0.060 | 0.30 | 0.10 |
| Secondary Outcomes | | | | | | | | |
| Stroke | 0.02 | 0.005 | 0.164 | 0.236 | 0.038 | 0.050 | 0.45 | 0.02 |
| Pulmonary embolism | 0.10 | 0.13 | 0.475 | 0.488 | −0.033 | 0.092 | 0.72 | 0.30 |
| Colorectal cancer | 0.44 | 0.44 | −0.256 | 0.343 | 0.109 | 0.080 | 0.17 | 0.29 |
| Hip fracture | 0.03 | 0.03 | −0.727 | 0.457 | 0.065 | 0.084 | 0.44 | 0.07 |
| All-cause mortality | 0.55 | 0.65 | 0.291 | 0.187 | −0.055 | 0.036 | 0.13 | 0.29 |

[1]Test for zero value restricted mean survival time (RMST) difference between groups.

[2]Test for proportional hazards (PH) by testing $\beta_2 = 0$ in HR model.

[3]Two degrees of freedom (DF) test $\beta_1 = \beta_2 = 0$ in HR model.

the two tests for most clinical outcomes for both hormone therapy trials. However, the RMST test does not identify ($p = 0.17$) an increased breast cancer risk in the active CEE+MPA group, whereas the logrank $p = 0.05$. This is an important difference because a breast cancer risk elevation was the trigger for the early stoppage of the CEE+MPA trial, which evidently impacted national breast cancer rates (Ravdin et al., 2007). The Table 2 $p$-values illustrate the importance of simultaneous consideration for the set of outcomes that plausibly differ between randomization groups, a topic that we will return to after discussing some properties of semiparametric hazard ratio (Cox) regression in this type of setting.

## 4. HAZARD RATIO MODELING AND ESTIMATION

### 4.1 Properties

A hazard ratio regression (Cox) model for the hazard rate at follow-up time $t$ can be written

$$\Lambda\{dt; Z(t)\} = \Lambda_0(dt)\exp\{x(t)\beta\}, \tag{7}$$

where $x(t) = \{x_1(t), x_2(t), \ldots, x_p(t)\}$ is a fixed length modeled (row) regression vector formed from $\{t, Z(t)\}$, $\beta = (\beta_1, \ldots, \beta_p)'$ is a corresponding (column) $p$-vector and $\Lambda_0(dt)$ is an unspecified 'baseline' hazard rate, at follow-up time $t$ and modeled regression vector value $x(t) \equiv (0, \ldots, 0)$. Note that (7) factors the hazard rates into a hazard ratio component $\exp\{x(t)\beta\}$ that models variations in hazard rates as a function of the regression vector $x(t)$ that may distinguish study subjects, and an absolute (baseline) hazard rate component $\Lambda_0$, where $\Lambda_0(t) = \int_0^t \Lambda_0(ds)$. The HR component characterizes dependence of the hazard rate on preceding covariate histories over time. For example, in a randomized, controlled trial one may consider $x(t) \equiv z$ where $z = 0$ and $z = 1$ again indicate control and active treatment assignment, respectively, thereby modeling a constant hazard ratio $e^{\beta}$ for the active compared to the control treatment groups. More generally, the HR may vary with follow-up time. Dependencies of this type can be modeled through 'defined' time-dependent covariate specifications, such as $x(t) = \{x_1(t), x_2(t)\} = \{zI(0 \le t < t_0), zI(t_0 \le t)\}$, where $I(\cdot)$ denotes an indicator function, which allows possibly distinct hazard ratios $e^{\beta_1}$ and $e^{\beta_2}$ according to whether follow-up time is $< t_0$ or $\ge t_0$; or $x(t) = (z, zt)$ which gives a treatment hazard ratio function $e^{\beta_1+\beta_2 t}$ that varies smoothly with follow-up time in an increasing ($\beta_2 > 0$), decreasing ($\beta_2 < 0$) or constant ($\beta_2 = 0$) fashion. In general HR modeling factors the hazard rates into a nonparametric and a parametric component. This

model admits a very simple and computationally reliable procedure for estimating the HR component parameter $\beta$ by solving the partial likelihood (Cox, 1975) score equation $U(\beta, \tau) = 0$ where

$$U(\beta, t) = \sum_{i=1}^{n} \int_0^t \{x_i(s) - E(s; \beta)\} N_i(ds), \tag{8}$$

based on a random sample $\{S_i = T_i \wedge C_i, \delta_i = I[T_i = S_i], Z_i(S_i), i = 1, \ldots, n\}$ from a study population.

Also in (8)

$$E(s; \beta) = \sum_{i=1}^{n} Y_i(s) x_i(s) e^{x_i(s)\beta} / \sum_{i=1}^{n} Y_i(s) e^{x_i(s)\beta}.$$

Under independent and identically distributed (IID) conditions for $\{S_i, \delta_i, Z_i(S_i)\}, i = 1, \ldots, n$, asymptotic distribution theory for $\hat{\beta}$ solving $U(\beta, \tau) = 0$ can be developed using martingale convergence theory in a manner that extends to counting process intensity modeling for multivariate outcomes on the same failure time axis (Andersen and Gill, 1982). Alternatively, asymptotic distribution theory for $\hat{\beta}$ can be developed using empirical process methods, in a manner that generalizes to models of the form (7) for marginal hazards with multivariate outcomes on the same or different failure time axes (e.g., Wei, Lin and Weissfeld, 1989, Spiekerman and Lin, 1998). Under IID assumptions and regularity conditions, the empirical process approach implies a mean zero asymptotic Gaussian distribution for $n^{1/2}(\hat{\beta} - \beta)$ with variance matrix that is consistently estimated by $n\hat{\Sigma}(\hat{\beta})^{-1}\hat{A}(\hat{\beta})\hat{\Sigma}(\hat{\beta})^{-1}$, where $\hat{\Sigma}(\hat{\beta}) = -\partial U(\hat{\beta}, t)/\partial\hat{\beta}'$, and

$$\hat{A}(\hat{\beta}) = \sum_{i=1}^{n} \left[ \int_0^{\tau} \{x_i(t) - E(t, \hat{\beta})\} \hat{M}_i(dt; \hat{\beta}) \right]^{\otimes 2},$$

with $\hat{M}_i(dt; \hat{\beta}) = N_i(dt) - Y_i(t) e^{x_i(t)\hat{\beta}} \hat{\Lambda}_0(dt; \hat{\beta})$, and $\hat{\Lambda}_0$ is the Breslow–Aalen estimator of the baseline hazard function $\Lambda_0$ given by

$$\hat{\Lambda}_0(dt; \beta) = \sum_{i=1}^{n} N_i(dt) / \sum_{i=1}^{n} Y_i(t) e^{x_i(t)\beta}. \tag{9}$$

Note that $n\hat{\Sigma}(\hat{\beta})^{-1}$ is the usual variance estimator for $n^{1/2}(\hat{\beta} - \beta)$ under model (7). These are results of considerable generality in view of the time-dependent feature of the parametric component of (7). For example the treatment hazard ratio in a randomized controlled trial can be modeled to allow distinct values over a partition of the follow-up period, embracing a class of models broader than overall proportional hazards. Furthermore, the baseline hazard function $\Lambda_0$ in (7) can be allowed to have distinct values over a fixed number of possibly time-dependent strata, defined from $\{t, Z(t)\}$ at follow-up time $t$, without adding appreciable complexity to parameter estimation.

The model (7) and its extensions separates comparative rates through the hazard ratio factor, from absolute rates that reflect also the baseline hazard rate function(s). Intuitively, comparative rates may be well modeled with few parameters, while absolute rates may depend on many details of the study population, as may reflect eligibility and exclusionary criteria in a clinical trial or cohort study setting.

### 4.2 Illustration

The nonparametric AHR and RMST difference ITT analyses of Figure 1 and Supplementary Figure 1 have some limitations. Specifically the estimators tend to be noisy, especially the AHR. Also neither estimator acknowledges the variation, from about 3.5 to 8.5 years, in the time from randomization to the end of an individual's intervention period. Also, it may be helpful to acknowledge the dependence of disease rates on participant age and other prominent risk factors.

To do so a Cox model (7) was applied with baseline hazard rate stratification on age at enrollment (50–54, 55–59, 60–69, 70–79), randomization status in the companion randomized low-fat dietary pattern trial (intervention, control, not randomized), prior history of the disease under analysis (if applicable), and study phase (time-dependent intervention and post-intervention). Table 1 shows logHR ($\beta$) estimators and estimated standard errors comparing the intervention and control groups for each of the global index outcomes, for both hormone therapy trials, along with a test of equality of phase-specific HRs for each outcome. Also shown are simple proportional hazards estimators and estimated 95% CIs over cumulative follow-up, along with corresponding estimators from an inverse variance weighted linear combination of logHR estimates from the two study phases. The two sets of estimators are essentially identical. Of course, HR estimates may vary also within study phases. Supplementary Table shows results of intervention phase ITT analyses that allow distinct HRs by follow-up year, for CHD and breast cancer. Also shown are intervention phase results under a simple proportional hazards assumption, which again are essentially identical to those under inverse variance weighting of follow-up year specific logHR estimates, for both outcomes and both trials. An informal description as to why such agreement can be expected is also included in Supplementary Material. See also O'Quigley (2008, pp. 226–7) for related arguments, and Murphy and Sen (1991) for formal developments.

In summary, although it is appropriate to criticize Cox modeling under an oversimplified proportional hazards assumption, the resulting HR estimator and confidence interval may often provide a good approximation to an HR estimator derived from an appropriately weighted linear combination of logHR estimators over a time axis partition. It is a drawback that the inverse variance weighting

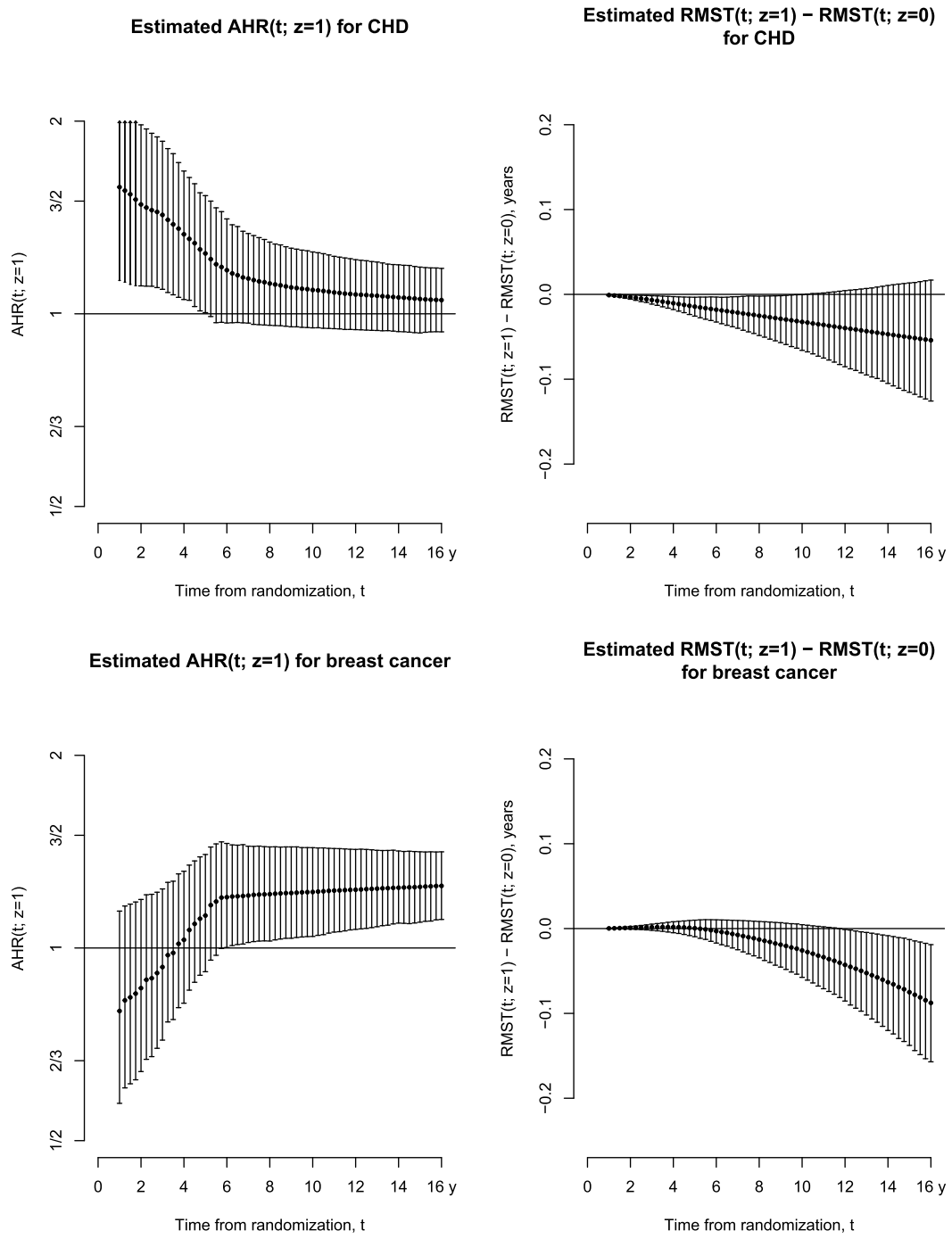


FIG. 2. *Semiparametrically estimated average hazard ratio* (*AHRs*) *and restricted mean survival time differences* (*RMST active-RMST, placebo*) *with nonparametric bootstrap* 95% *confidence intervals* (1000 *bootstrap samples*) *from the Women's Health Initiative estrogen plus progestin* (*CEE+MPA*) *trial among* 16,608 *postmenopausal US women with uterus, under a hazard rate model defined by expressions* (7) *and* (10).

may depend on censoring rates, though in the WHI trial setting with mostly administrative censoring, this is unlikely to be an important issue.

Importantly, with careful modeling of HR dependencies on time, one can expect to estimate the ITT HR function well, and to do so also for related functions, such as AHRs and RMST differences. For example, Figure 2 and Supplementary Figure 2 show estimators and pointwise 95% confidence intervals (1000 bootstrap samples) like those in Figure 1 and Supplementary Figure 1 under a Cox model (1) with

$$x_i(t) = \{zI(t \le t_{0i}), ztI(t \le t_{0i}), zI(t > t_{0i}), z(t - t_{0i})I(t > t_{0i})\} \tag{10}$$

where $t_{0i}$ denotes time from randomization to the end of the trial intervention period for the $i$th participant. This model allows separate HR function estimation in the intervention and post-intervention trial phases, while

allowing a smooth form of departure from proportionality in each phase. Estimated average HR functions, $\{\hat{\text{A}}\text{HR}(t; z, t_0), t \geq 0\}$, for the primary CHD and breast cancer outcomes for a participant having a $t_{0i} = 5.6$ year intervention period are shown over cumulative follow-up for the CEE+MPA trial on the left-hand side of Figure 2. Under the assumed HR form each point $(t)$ on these plots provides a consistent estimator of average HR from randomization to $t$. For CHD in the CEE+MPA trial one obtains maximum partial likelihood estimators (estimated standard errors) for the four HR parameters of 0.516 (0.198), −0.119 (0.049), 0.011 (0.113) and −0.001 (0.014) respectively. The resulting AHR estimator shows a risk elevation over early follow-up periods, which dissipates over post-randomization follow-up. For example, over the first 2, 6 and 16 years $\hat{\text{A}}\text{HR}$ (95% CI) estimates are 1.48 (1.11, 1.97), 1.17 (0.97, 1.41) and 1.05 (0.94, 1.18), respectively. Corresponding RMST difference (95% CI) estimates are difficult to read from Figure 2 (right side) over early follow-up periods, but convey a similar message with corresponding values −0.003 (−0.006, −0.001), −0.018 (−0.033, −0.003) and −0.054 (−0.126, 0.017) respectively.

For breast cancer in the CEE+MPA trial the four regression parameter estimates (95% CIs) from partial likelihood maximization are −0.318 (0.214), 0.156 (0.050), 0.221 (0.135) and 0.005 (0.018). From these one sees a nonsignificantly reduced AHR over early follow-up periods that changes to an elevated AHR over the later intervention period, and that remains elevated and fairly constant over long-term follow-up. For example, AHR (95%CI) estimates over 2, 6 and 16 year periods following randomization are 0.87 (0.64, 1.19), 1.20 (1.00, 1.46) and 1.25 (1.11, 1.41). Corresponding estimates (95% CIs) for the RMST difference convey a somewhat different message with values of 0.001 (−0.001, 0.003), −0.003 (−0.017, 0.010), −.088 (−0.157, −0.019) at 2, 6, and 16 years respectively, with evidence of reduction in RMST in the intervention group not arising until about 12 years of cumulative intervention and follow-up. Once again, since breast cancer risk elevation was the principal trigger for the early stoppage of the CEE+MPA intervention following 5.6 year (median) follow-up period, this is a practically important difference between these two summary measures.

For clinical interpretation it would be accurate to say that postmenopausal US women, like those enrolled in the WHI trial, who are assigned to CEE+MPA with a 5.6 year intervention period experience about an average 20% increase in breast cancer incidence over 6 years of intervention and follow-up, compared to similar women assigned to placebo. A risk elevation of about this same magnitude is maintained over a 16 year cumulative intervention and follow-up period. From the RMST function estimator one could summarize that comparison of the same two populations of women yielded a breast cancer–free survival time that was about 4.6 weeks (0.088 years) shorter with CEE+MPA assignment over 16 years of intervention and follow-up, though a reduction was not clearly evident until about 12 years of cumulative follow-up.

Similarly for CHD, from the AHR estimator one could summarize in the population of postmenopausal US women studied, that assignment to 5.6 years of CEE+MPA led to about a 50% increase in CHD incidence over the first two years, but this adverse effect dissipated over long-term intervention and follow-up. From the RMST estimator, one could summarize that the population of postmenopausal US women have an estimated reduction by 1.1 days (0.003 years) in disease free 'survival' time over the first two years of intervention if assigned to CEE+MPA, but this small reduction dissipated over a longer-term period of intervention and follow-up. Even though breast cancer and CHD are relatively common outcomes over the lifespan of US women, the small incidence rates experienced over a follow-up period of a few years in prevention trials or cohort studies would appear to reduce the utility of the RMST contrasts as a clinical communication tool.

Supplementary Figure 2 presents corresponding analyses under the same underlying hazard ratio model for the CEE trial, now for a participant having 7.2 years from randomization to the end of her intervention period. There is no clear evidence of a CHD influence using either AHR or RMST difference to summarize the data. A weak unfavorable trend over early time periods is followed by a weak favorable trend over a longer period of intervention and follow-up, but neither trend is close to significant. For breast cancer, the AHR assessment provides some evidence for an early benefit, which becomes stronger over a long period of intervention and follow-up. The RMST difference also provides evidence of breast cancer benefit over long cumulative follow-up periods.

For null hypothesis testing one could consider $p$-values for a simultaneous test of zero values for the four coefficients of the regression vector (10). Alternatively, one could consider a simpler test of $\beta_1 = \beta_2 = 0$ with a Cox model having $x(t) = (z, zt)$. As shown in Table 2 this test has $p$-value 0.004 for breast cancer, with a test of overall proportional hazards ($\beta_2 = 0$) having $p = 0.008$. For comparability with other null hypothesis tests, baseline hazard stratification is dropped from these Cox model–based tests. Note that $p$-values for testing $\beta_1 = \beta_2 = 0$ tend to be as extreme or more extreme than those for the other tests shown, even when there is little evidence against overall proportionality ($\beta_2 = 0$).

## 5. MULTIVARIATE FAILURE TIME ITT ANALYSES

As illustrated in Table 1 a more complete view of ITT effects in randomized trials can be obtained by consider-

ing multiple time-to-response outcomes that may plausibly be influenced by the intervention. For example, hazard ratio models (7) for marginal hazard rates for each outcome have well developed distribution theory for parameter estimates based on (8) and (9) for each outcome (e.g., Spiekerman and Lin, 1998).

In settings such as the WHI trials with outcomes that are rare during the study follow-up period, most pertinent information beyond that for marginal single failure hazard estimands derives from marginal dual outcome hazard rate analyses. Consider failure time variates $(T_1, T_2)$ subject to right censoring by variate $(C_1, C_2)$, and a covariate process $Z$ that may be two dimensional. Let $Z(t_1, t_2)$ denote the history of $Z$ prior to $T_1 < t_1$ and $T_2 < t_2$. One can define the marginal dual outcome hazard rate at follow-up time $t_1$ for $T_1$ and $t_2$ for $T_2$, given $Z$, by $\Lambda\{dt_1, dt_2; Z(t_1, t_2)\} = P\{t_1 \le T_1 < t_1 + dt_1, t_2 \le T_2 < t_2 + dt_2; T_1 \ge t_1, T_2 \ge t_2, Z(t_1, t_2)\}$. In the randomized trial setting, with $Z(t_1, t_2) \equiv z$ a corresponding dual outcome hazard function can be estimated nonparametrically in each treatment group for additional causal treatment comparisons. One can also define a two-dimensional average dual outcome hazard function estimand by

$$\begin{aligned}&\mathrm{AHR}(t_1, t_2)\\ &= \int_0^{t_1}\int_0^{t_2} \frac{\Lambda\{ds_1, ds_2; z=1\}}{\Lambda\{ds_1, ds_2; z=0)\}} F(ds_1, ds_2; z=0)\\ &\quad /\{1 - F(t_1, 0; z=0) - F(0, t_2; z=0)\\ &\quad + F(t_1, t_2; z=0)\},\end{aligned}$$

where $F(t_1, t_2; z)$ is the joint survival function for $(T_1, T_2)$ in randomization group $z$. This estimand can be readily estimated nonparametrically under IID conditions, for example, using the Volterra estimator of $F$ (e.g., Gill, van der Laan and Wellner, 1995, Prentice and Zhao, 2019, p. 55) and corresponding asymptotic distributions can be derived using empirical process theory, and bootstrap procedures can be applied for confidence interval and confidence band calculation.

Alternatively one can specify a Cox-type model for dual outcome hazard rates according to

$$\Lambda\{dt_1, dt_2; Z(t_1, t_2)\} = \Lambda(dt_1, dt_2)\exp\{x(t_1, t_2)\beta\},$$

where $\Lambda(dt_1, dt_2)$ is a 'baseline' dual outcome hazard rate at $(t_1, t_2)$ and value $x(t_1, t_2) \equiv 0$ for fixed length (row) vector $x$, with $x(t_1, t_2)$ defined as a function of $\{t_1, t_2, Z(t_1, t_2)\}$, while column vector $\beta$ is a corresponding dual outcome hazard ratio parameter to be estimated. Empirical process theory leads to asymptotic results for marginal single and dual outcome hazard rate parameters jointly, and bootstrap procedures are applicable for related estimation (Prentice and Zhao, 2020).

### 5.1 Illustration

Table 3 shows numbers of dual outcomes in the intervention and control groups in the WHI hormone therapy trials, along with dual outcome (proportional) hazard ratio estimates and 95% CIs, for the primary CHD and breast cancer outcomes, and for all-cause mortality, over cumulative follow-up. For example, in the CEE+MPA trial, one sees evidence of risk elevation for the dual outcome of breast cancer followed by death from any cause. This type of analysis adds information on the causal effects of being assigned to active CEE+MPA. Of course, a simple proportional hazards specification for these dual outcome hazard rates is likely an oversimplification, but the resulting estimated hazard ratios can presumably be thought of as approximately arising from inverse variance weighted linear combination of log-HR estimates over a partition of the two-dimensional follow-up regions. See Prentice et al. (2020b) for similar analyses for additional clinical outcomes, and for analyses that exercise the time-varying hazard ratio feature to distinguish HRs according to which of two failure times occurred first. These types of Cox model analyses can be powerful also for extending ITT comparisons to include studies of disease pathways and mechanisms.

## 6. DISCUSSION

Throughout this presentation, we have emphasized hazard function contrasts, since these derive a causal interpretation from the independence of randomization assignments and pre-randomizaiton risk factors for time-to-response outcomes. The purity of these contrasts encourage their broad use in the reporting of randomized trials,

TABLE 3

*Dual outcome cases, hazard ratio (HR) estimates and 95% confidence intervals (CIs) over cumulative follow-up in the WHI hormone therapy trials*

| | | CEE+MPA Trial | | | | CEE Trial | | | |
|---|---|---|---|---|---|---|---|---|---|
| Outcome/Outcomes | | CHD | | Breast Cancer | | CHD | | Breast Cancer | |
| Breast Cancer | Cases I[1] Cases C[1] | 41 | 41 | | | 25 | 34 | | |
| | HR (95% CI) | 0.96 (1.03, 1.48) | | | | 0.73 (0.43, 1.24) | | | |
| Death | Cases I Cases C | 447 | 427 | 189 | 145 | 348 | 379 | 90 | 113 |
| | HR (95% CI) | 1.03 (0.90, 1.17) | | 1.26 (1.01, 1.56) | | 0.93 (0.80, 1.08) | | 0.80 (0.61, 1.06) | |

[1]I—intervention group; C—control group.

including analyses that aim to discover the temporal patterns and biological bases for the treatment effects under study. We think that hazard ratios provide an important focus for ITT comparisons, and that the flexibilties and reliable computations associated with the Cox model help to justify its central place in the reporting of a range of ITT analyses. While we acknowledge that it may be complex to accurately describe hazard functions and hazard ratios to nonstatistical audiences, statements such as, 'the intervention group had an elevated risk for a certain outcome by about 30% over the first few years following randomization' are readily accepted by collaborators coming from various disciplines, in our experience. Of course, the identification of the preferred methods for communicating ITT results is an important research goal in itself. For example RMST contrasts may be useful for this purpose for common outcomes, while relative statements may be preferred for rare outcomes. Note that even in settings where ratio measures, such as HRs, provide the principal reported contrasts, additional information comparing absolute risks should also be reported to provide context for overall benefit versus risk considerations.

Along with our reliance on ITT analysis to the extent possible, we are quick to acknowledge the wealth of additional questions of interest that typically attend the reporting of a randomized, controlled trial. These are 'what if' questions: For example, 'What would the treatment effects be if study subjects fully adhered to the regimen under study?' 'What would the hazard ratio be at follow-up time $t$ if there had not been differential selection due to earlier outcomes, either for the study outcome under consideration or for competing outcomes?' 'How would the trial results look if study subjects had not made changes beyond those specified in the protocol?'. These are important questions, but our ability to address them requires the identification, measurement, and proper modeling of post-randomization variables, $L$, that are relevant to these topics, and related analyses conditional on $L$ become observational in nature. As Hernán and Robins (2020, p. 29) write, 'Unfortunately, no matter how many variables are included in $L$, there is no way to test that the assumption is correct, which makes causal inference from observational data a risky task'. The assumption in question is that of conditional exchangeability whereby treatment assignment is orthogonal to outcome given $L$. These considerations should not dissuade one from attempting to make treatment contrasts in randomized trials that, for example, adjust from nonadherence or for post-randomization confounding, but these analyses do not enjoy the same reliability as do ITT comparisons, and related causality claims are necessarily attended by assumptions that typically cannot be fully verified. Hence, causal inference via ITT comparisons may be distinguished by validity in conjunction with some limitations on interpretation, whereas causal inference via emulation of randomized trials may target parameters of clearer interpretation but lack the validity of ITT comparisons.

It seems to us that the theory of causal inference via potential outcomes and conditional exchangeability provides an important pathway to inference on parameters that can enhance the overall interpretation and impact of randomized trials. We expect that hazard functions, and hazard ratio modeling, can provide useful foundations upon which to continue the development of these procedures (see also Hernán and Robins, 2020, p. 211).

## FUNDING

This work was partially supported by National Institutes of Health awards HHSN268201100046C and P30CA015704.

## SUPPLEMENTARY MATERIAL

**Supplement to "Intention-to-Treat Comparisons in Randomized Trials"** (DOI: 10.1214/21-STS830SUPP; .pdf). Supplementary information.

## REFERENCES

AALEN, O., BORGAN, Ø. and GJESSING, H. (2010). *Survival and Event History Analysis*: *A Process Point of View*. Springer, New York.

AALEN, O. O., COOK, R. J. and RØYSLAND, K. (2015). Does Cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime Data Anal*. **21** 579–593. MR3397507 https://doi.org/10.1007/s10985-015-9335-y

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist*. **10** 1100–1120. MR0673646

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. *Springer Series in Statistics*. Springer, New York. MR1198884 https://doi.org/10.1007/978-1-4612-4348-9

ANDERSON, G., LIMACHER, M., ASSAF, A., BASSFORD, T., BERESFORD, S., BLACK, H., BONDS, D., BRUNNER, R., BRZYSKI, R. et al. (2004). Effects of conjugated equine estrogen in postmenopausal women with hysterectomy: The Women's Health Initiative randomized controlled trial. *JAMA* **291** 1701–1712.

BEGUN, J. M., HALL, W. J., HUANG, W.-M. and WELLNER, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist*. **11** 432–452. MR0696057 https://doi.org/10.1214/aos/1176346151

CHLEBOWSKI, R. T., KULLER, L. H., PRENTICE, R. L., STEFANICK, M. L., MANSON, J. E., GASS, M., ARAGAKI, A. K., OCKENE, J. K., LANE, D. S. et al. (2009). Breast cancer after use of estrogen plus progestin in postmenopausal women. *N. Engl. J. Med*. **360** 573–587.

CHLEBOWSKI, R. T., ANDERSON, G. L., ARAGAKI, A. K., MANSON, J. E., STEFANICK, M. L., PAN, K., BARRINGTON, W., KULLER, L. H., SIMON, M. S. et al. (2020). Association of menopausal hormone therapy with breast cancer incidence and mortality during long-term follow-up of the women's health initiative randomized clinical trials. *JAMA* **324** 369–380. https://doi.org/10.1001/jama.2020.9482

COOK, R. J. and LAWLESS, J. F. (2018). *Multistate Models for the Analysis of Life History Data*. *Monographs on Statistics and Applied Probability* **158**. CRC Press, Boca Raton, FL. MR3838371 https://doi.org/10.1201/9781315119731

COX, D. R. (1972). Regression models and life-tables(with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276. MR0400509 https://doi.org/10.1093/biomet/62.2.269

COX, D. R. (1992). Causality: Some statistical aspects. *J. Roy. Statist. Soc. Ser. A* **155** 291–301. MR1157712 https://doi.org/10.2307/2982962

FEIGL, P. and ZELEN, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* 826–838.

FINE, J. P., JIANG, H. and CHAPPELL, R. (2001). On semi-competing risks data. *Biometrika* **88** 907–919. MR1872209 https://doi.org/10.1093/biomet/88.4.907

GILL, R. D., VAN DER LAAN, M. J. and WELLNER, J. A. (1995). Inefficient estimators of the bivariate survival function for three models. *Ann. Inst. Henri Poincaré Probab. Stat.* **31** 545–597. MR1338452

HERNÁN, M. A. (2010). The hazards of hazard ratios. *Epidemiology* **21** 13–15.

HERNÁN, M. A. and ROBINS, J. M. (2020). *Causal Inference*: *What If*. CRC Press/CRC, Boca Raton, FL.

HSIA, J., LANGER, R., MANSON, J., KULLER, L., JOHNSON, K., HENDRIX, S., PETTINGER, M., HECKBERT, S., GREEP, N. et al. (2006). Women's health initiative investigators. Conjugated equine estrogens and coronary heart disease: The Women's Health Initiative. *Arch. Intern. Med.* **166** 357–365.

KALBFLEISCH, J. D. and PRENTICE, R. L. (1981). Estimation of the average hazard ratio. *Biometrika* **68** 105–112. MR0614947 https://doi.org/10.1093/biomet/68.1.105

KALBFLEISCH, J. D. and PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed. *Wiley Series in Probability and Statistics*. Wiley Interscience, Hoboken, NJ. MR1924807 https://doi.org/10.1002/9781118032985

LAGAKOS, S. W. and SCHOENFELD, D. A. (1984). Properties of proportional-hazards score tests under misspecified regression models. *Biometrics* **40** 1037–1048. MR0786178 https://doi.org/10.2307/2531154

MANSON, J. E., HSIA, J., JOHNSON, K. C., ROSSOUW, J. E., ASSAF, A. R., LASSER, N. L., TREVISAN, M., BLACK, H. R., HECKBERT, S. R. et al. (2003). Estrogen plus progestin and the risk of coronary heart disease. *N. Engl. J. Med.* **349** 523–534.

MANSON, J. E., CHLEBOWSKI, R. T., STEFANICK, M. L., ARAGAKI, A. K., ROSSOUW, J. E., PRENTICE, R. L., ANDERSON, G., HOWARD, B. V., THOMSON, C. A. et al. (2013). Menopausal hormone therapy and health outcomes during the e intervention and extended poststopping phases of the Women's Health Initiative randomized trials. *J. Am. Med. Assoc.* **310** 1353–1368.

MANSON, J. E., ARAGAKI, A. K., BASSUK, S. S., CHLEBOWSKI, R. T., ANDERSON, G. L., ROSSOUW, J. E., HOWARD, B. V., THOMSON, C. A., STEFANICK, M. L. et al. (2019). Menopausal estrogen-alone therapy and health outcomes in women with and without bilateral oophorectomy: A randomized trial. *Ann. Intern. Med.* **171** 406–414. https://doi.org/10.7326/M19-0274

MANTEL, N. and HAENZEL, W. (1959). Statistical analysis of data from retrospective studies. *J. Natl. Cancer Inst.* **22** 718–748.

MARTINUSSEN, T., VANSTEELANDT, S. and ANDERSEN, P. K. (2020). Subtleties in the interpretation of hazard contrasts. *Lifetime Data Anal.* **26** 833–855. MR4148449 https://doi.org/10.1007/s10985-020-09501-5

MURPHY, S. A. and SEN, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Process. Appl.* **39** 153–180. MR1135092 https://doi.org/10.1016/0304-4149(91)90039-F

O'QUIGLEY, J. (2008). *Proportional Hazards Regression*. *Statistics for Biology and Health*. Springer, New York. MR2400249 https://doi.org/10.1007/978-0-387-68639-4

PRENTICE, R. L. and ZHAO, S. (2019). *The Statistical Analysis of Multivariate Failure Time Data*: *A Marginal Modeling Approach*. *Monographs on Statistics and Applied Probability* **163**. CRC Press, Boca Raton, FL. MR3966434 https://doi.org/10.1201/9780429162367

PRENTICE, R. L. and ZHAO, S. (2020). Regression models and multivariate life tables. *J. Amer. Statist. Assoc.* https://doi.org/10.1080/01621459.2020.1713792

PRENTICE, R. L., LANGER, R., STEFANICK, M. L., HOWARD, B. V., PETTINGER, M., ANDERSON, G., BARAD, D., CURB, J. D., KOTCHEN, J. et al. (2005). Combined postmenopausal hormone therapy and cardiovascular disease: Toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. *Am. J. Epidemiol.* **162** 404–414.

PRENTICE, R. L., LANGER, R. D., STEFANICK, M. L., HOWARD, B. V., PETTINGER, M., ANDERSON, G. L., BARAD, D., CURB, J. D., KOTCHEN, J. et al. (2006). Combined analysis of women's health initiative observational and clinical trial data on postmenopausal hormone treatment and cardiovascular disease. *Am. J. Epidemiol.* **163** 589–599.

PRENTICE, R. L., MANSON, J. E., LANGER, R. D., ANDERSON, G. L., PETTINGER, M., JACKSON, R. D., JOHNSON, K. C., KULLER, L. H., LANE, D. S. et al. (2009). Benefits and risks of postmenopausal hormone therapy when it is initiated soon after menopause. *Am. J. Epidemiol.* **170** 12–23. https://doi.org/10.1093/aje/kwp115

PRENTICE, R. L., ARAGAKI, A. K., CHLEBOWSKI, R. T., ROSSOUW, J. E., ANDERSON, G. L., STEFANICK, M. L., WACTAWSKI-WENDE, J., KULLER, L. H., WALLACE, R. et al. (2020). Randomized trial evaluation of the benefits and risks of menopausal hormone therapy among women 50–59 years of age. *Am. J. Epidemiol.* **190** 365–375.

PRENTICE, R. L., ARAGAKI, A. K., CHLEBOWSKI, R. T., ZHAO, S., ANDERSON, G. L., ROSSOUW, J. E., WALLACE, R., BANACK, H., SHADYAB, A. H. et al. (2020b). Dual outcome intention-to-treat analyses in the women's health initiative randomized controlled hormone therapy trials. *Am. J. Epidemiol.* **189** 972–981.

PRENTICE, R. L and ARAGAKI, A. K (2022). Supplement to "Intention-To-Treat Comparisons in Randomized Trials." https://doi.org/10.1214/21-STS830SUPP

RAVDIN, P. M., CRONIN, K. A., HOWLADER, N., BERG, C. D., CHLEBOWSKI, R. T., FEUER, E. J., EDWARDS, B. K. and BERRY, D. A. (2007). The decrease in breast-cancer incidence in 2003 in the United States. *N. Engl. J. Med.* **356** 1670–1674.

ROBINS, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512. (Errata in Computers and Mathematics with Applications 1987; 14:917–921).

ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23** 2379–2412. MR1293185 https://doi.org/10.1080/03610929408831393

ROSSOUW, J., ANDERSON, G., PRENTICE, R., LACROIX, A., KOOPERBERG, C., STEFANICK, M., JACKSON, R., BERESFORD, S., HOWARD, B. et al. (2002). Risks and benefits of estrogen plus progestin in healthy postmenopausal women: Principal

results from the Women's Health Initiative randomized controlled trial. *J. Am. Med. Assoc.* **288** 321–333.

ROTH, J. A., ETZIONI, R., WATERS, T. M., PETTINGER, M., ROSSOUW, J. E., ANDERSON, G. L., CHLEBOWSKI, R. T., MANSON, J. E., HLATKY, M. et al. (2014). Economic return from the Women's Health Initiative estrogen plus progestin clinical trial: A modeling study. *Ann. Intern. Med.* **160** 594–602. https://doi.org/10.7326/M13-2348

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

SCHEMPER, M., WAKOUNIG, S. and HEINZE, G. (2009). The estimation of average hazard ratios by weighted Cox regression. *Stat. Med.* **28** 2473–2489. MR2751535 https://doi.org/10.1002/sim.3623

SPIEKERMAN, C. F. and LIN, D. Y. (1998). Marginal regression models for multivariate failure time data. *J. Amer. Statist. Assoc.* **93** 1164–1175. MR1649210 https://doi.org/10.2307/2669859

STRUTHERS, C. A. and KALBFLEISCH, J. D. (1986). Misspecified proportional hazard models. *Biometrika* **73** 363–369. MR0855896 https://doi.org/10.1093/biomet/73.2.363

UNO, H., WITTES, J., FU, H., SOLOMON, S. D., CLAGGETT, B., TIAN, L., CAI, T., PFEFFER, M. A., EVANS, S. R. et al. (2015). Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Ann. Intern. Med.* **163** 127–134.

WEI, L. J., LIN, D. Y. and WEISSFELD, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Statist. Assoc.* **84** 1065–1073. MR1134494

ZHAO, L., CLAGGETT, B., TIAN, L., UNO, H., PFEFFER, M. A., SOLOMON, S. D., TRIPPA, L. and WEI, L. J. (2016). On the restricted mean survival time curve in survival analysis. *Biometrics* **72** 215–221. MR3500590 https://doi.org/10.1111/biom.12384