CLINICAL TRIALS

Clinical Trials
2018, Vol. 15(2) 178–188

Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1740774518759281
journals.sagepub.com/home/ctj
SAGE

Restricted mean survival time: Does covariate adjustment improve precision in randomized clinical trials?

Theodore Karrison[1] and Masha Kocherginsky[2]

Abstract

Background: Restricted mean survival time is a measure of average survival time up to a specified time point. There has been an increased interest in using restricted mean survival time to compare treatment arms in randomized clinical trials because such comparisons do not rely on proportional hazards or other assumptions about the nature of the relationship between survival curves.

Methods: This article addresses the question of whether covariate adjustment in randomized clinical trials that compare restricted mean survival times improves precision of the estimated treatment effect (difference in restricted mean survival times between treatment arms). Although precision generally increases in linear models when prognostic covariates are added, this is not necessarily the case in non-linear models. For example, in logistic and Cox regression, the standard error of the estimated treatment effect does not decrease when prognostic covariates are added, although the situation is complicated in those settings because the estimand changes as well. Because estimation of restricted mean survival time in the manner described in this article is also based on a model that is non-linear in the covariates, we investigate whether the comparison of restricted mean survival times with adjustment for covariates leads to a reduction in the standard error of the estimated treatment effect relative to the unadjusted estimator or whether covariate adjustment provides no improvement in precision. Chen and Tsiatis suggest that precision will increase if covariates are chosen judiciously. We present results of simulation studies that compare unadjusted versus adjusted comparisons of restricted mean survival time between treatment arms in randomized clinical trials.

Results: We find that for comparison of restricted means in a randomized clinical trial, adjusting for covariates that are associated with survival increases precision and therefore statistical power, relative to the unadjusted estimator. Omitting important covariates results in less precision but estimates remain unbiased.

Conclusion: When comparing restricted means in a randomized clinical trial, adjusting for prognostic covariates can improve precision and increase power.

Keywords

Restricted mean, covariate adjustment, efficiency, power

Introduction

The log-rank test and the Cox[1] proportional hazards regression model are two of the most popular procedures for comparing survival times in different treatment arms of a randomized clinical trial (RCT). The log-rank test is known to be most powerful under proportional hazards alternatives, and proportional hazards are usually assumed when fitting the Cox regression model, although the model can be extended to accommodate non-proportional hazards as described in Cox's original manuscript and by others.[2] A variety of parametric models are also available for analyzing survival data,[3] and these models can be applied under proportional hazards, accelerated failure time, and other frameworks.[4] All of these methods allow for censoring, that is, observations in which the event of

[1] Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA
[2] Department of Preventive Medicine, Northwestern University, Chicago, IL, USA

Corresponding author:
Theodore Karrison, Department of Public Health Sciences, The University of Chicago, 5841 South Maryland Avenue, MC2000, Chicago, IL 60637, USA.
Email: tkarrison@health.bsd.uchicago.edu

interest has not yet occurred, a common feature of survival data from RCTs.

Were it not for censoring, mean survival times could be compared in different groups using standard methods, such as two-sample *t*-tests or ordinary least-squares multiple regression if covariate effects or adjustment are of interest. Methods for least squares regression with censored data have been proposed,[5] but are rarely used because of computational issues and, until relatively recently, the lack of available software.[6] In addition to evaluating covariate effects on outcome, covariates are often included in an analysis for two other purposes: (1) to adjust the treatment effect for imbalances in prognostic factors between the treatment arms, although in an RCT this would be adjustment for random imbalances only and (2) to improve the precision of the estimated treatment effect by accounting for other sources of variation.

As censoring generally precludes estimation of the mean survival time, Irwin[7] proposed, as an alternative, to estimate the expectation of life limited (restricted) to a suitably chosen time t^*. Since survival time is a positive random variable, the mean μ could be obtained by integrating under the survival curve

$$\mu = \int_0^\infty S(t)dt \qquad (1)$$

where $S(t)$ is the survival function for the random variable $T>0$. Although one might consider substituting the Kaplan and Meier[8] estimate of the survival function into equation (1) to estimate μ, when censoring is heavy, $S(t)$ is often ill-determined, or even undefined, beyond a certain range. Instead, the restricted mean survival time (RMST) is given by

$$\mu_{t^*} = \int_0^{t^*} S(t)dt \qquad (2)$$

Thus, μ_{t^*} is simply the area under the survival curve up to the point of restriction t^*. μ_{t^*} can also be viewed as the mean of a new random variable taking the value T if $T<t^*$ and t^* if $T \geq t^*$, that is

$$\mu_{t^*} = E[\min(T, t^*)]$$

Kaplan and Meier discussed estimation of the restricted mean by substituting the product-limit estimator into equation (2), and Meier[9] established its asymptotic normality.

Why consider RMST? In addition to its simple interpretation as the mean "up to" time t^*, an important advantage is that comparison of RMSTs between two groups requires no assumptions about the relationship between the two survival curves. Royston and Parmar,[10] in consideration of violations of the proportional hazards assumption, concluded that "the hazard ratio cannot be recommended as a general measure of the treatment effect in a randomized controlled trial. [...] Restricted mean survival time may provide a practical way forward and deserves greater attention." Uno et al.[11] considered various alternatives to the hazard ratio, including RMST, for quantifying the difference between two survival curves and concluded that the difference or ratio of RMSTs provides a good summary measure. Trinquart et al.[12] conducted a meta-analysis of 54 randomized oncology trials in which they computed both the hazard ratio and the difference in RMSTs and remarked that "Our analysis also highlights how the difference in RMST provides a clinically meaningful summary of evidence. It allows for quantifying the absolute survival difference and grading the magnitude of clinical benefit." Chappell and Zhu[13] point out that, like means, RMSTs have the useful property of being additive, and Uno et al.[14] discuss the advantages of comparing RMSTs in non-inferiority studies with low event rates.

However, one should not lose sight of the fact that the restricted mean is just that—a *restricted* mean: it ignores everything beyond the point of restriction and cuts off the distribution at t^*. If the true survival curves remain separated beyond the point of restriction, the difference in restricted means will increase with t^*. Consequently, estimated differences can appear somewhat small and alternative effect measures, such as the ratio of RMSTs and expressing RMST as the percentage of potential life-years achieved,[15] can be helpful in the assessment of clinical benefit.

Karrison[15] and Royston and Parmar[10] provided sample size formulae for designing a clinical trial based on RMST and made recommendations for choosing the point of restriction t^*. Karrison[15] also compared the power of RMST with the log-rank test (asymptotically equivalent to the scores test from the Cox regression model) and the generalized Wilcoxon test under proportional hazards and non-proportional hazards alternatives. He found that RMST can provide increased power for early difference alternatives without sacrificing too much power relative to the log-rank test when proportional hazards hold. For late difference alternatives, however, the comparison of RMSTs entailed a loss in power relative to the log-rank test.

Covariate adjustment in linear and non-linear models

In linear models, adjusting for covariates that are associated with outcome increases the precision of the treatment effect estimator. In the classic analysis of covariance (ANCOVA) model, for example, this is achieved through a reduction in the residual variance. In non-randomized studies, adjustment for covariates is almost always necessary in order to reduce confounding, whereas, as mentioned above, in RCTs it serves the dual purpose of adjusting for random imbalances in prognostic factors between treatment arms, as well as of potentially improving precision.

A gain in precision, however, cannot be taken for granted in non-linear models. Robinson and Jewell[16] showed that in logistic regression, adjustment for covariates leads to a loss in precision (or at best no gain). Similarly, Ford et al.[17] demonstrated that in the Cox regression model, adjustment for prognostic covariates does not improve the precision of the estimated treatment effect. Further complicating the decision of whether to adjust for covariates is that omitting influential covariates in both logistic and Cox regression produces a treatment effect estimate that is "biased" toward the null. If the model is misspecified, tests of the null hypothesis are valid, but if the alternative hypothesis is true, the "bias" toward the null results in diminished power. Therefore, it is still beneficial to adjust for prognostic covariates.[18,19] Schoenfeld and Borenstein[20] provide an algorithm for calculating the power for logistic and proportional hazards models that incorporate covariates. As in Hauck et al.,[19] we have placed the term "bias" above in quotation marks because in the case of logistic and Cox regression, the estimand changes as covariates are added, and the unadjusted and adjusted models actually estimate different measures of treatment effect. Heuristically, Hauck et al. describe this as moving from a "population-averaged" interpretation for unadjusted estimates toward a more "subject-specific" effect in covariate-adjusted models, where covariates can be thought of as representing the subject effect.

As the model for estimating RMST as developed here is also non-linear in the covariates, we address the following questions in this article. How does comparison of restricted means between treatments in an RCT fare in regard to covariate adjustment: is it necessary to adjust for covariates to obtain an unbiased estimate of the treatment effect, and does adjustment for covariates improve precision and/or statistical power?

Methodology

Let $t_1^0, t_2^0, \ldots, t_n^0$ denote the true survival times from a sample of size n. The observed survival time for the ith individual is $t_i = \min(t_i^0, c_i)$ where c_i is the ith individual's censoring time. Let Δ_i be the indicator variable taking the value 1 if t_i corresponds to an event and the value 0 if t_i is a censored observation. Let g_i denote treatment group ($g_i = 1, 2$) and $\mathbf{z}_i$ a vector of covariates, so that the data consist of $(t_i, \Delta_i, g_i, \mathbf{z}_i)$ $i = 1, 2, \ldots, n$. Karrison[21] incorporated covariates into the analysis of RMST by fitting a piecewise exponential model

$$\lambda_g(t|\mathbf{z}) = \lambda_{gk}\exp(\boldsymbol{\beta}'\mathbf{z}), t \epsilon (l_{k-1}, l_k], g = 1, 2$$

where $\lambda_g(t|\mathbf{z})$ is the hazard rate at time t for an individual in treatment group g with covariate vector $\mathbf{z}$, and the time axis is divided into intervals $(0, l_1], (l_1, l_2], \ldots, (l_{K-1}, t^*]$. The key features of this model are that (1) covariates are assumed to have proportional hazards effects, whereas (2) the different underlying piecewise constant hazard functions in the two treatment arms avoids the proportional hazards assumption with respect to the treatment effect. The fact that the β coefficients are assumed to be the same for both treatment groups makes the model analogous to the standard ANCOVA model in this regard. Zucker[22] avoided the arbitrary specification of intervals and developed asymptotic theory for estimating RMST under the stratified Cox model

$$\lambda_g(t|\mathbf{z}) = \lambda_{0g}(t)\exp(\beta'\mathbf{z}) \tag{3}$$

where the baseline hazard function for group g, $\lambda_{0g}(t)$, is left completely unspecified. We will use Zucker's model (3) in what follows.

Zucker used the Breslow estimator for the cumulative underlying hazard function in group g

$$\hat{\Lambda}_{0g}(t) = \sum_{T_{(i)} \le t}\left[\sum_{j=1}^{n_g} Y_{gj}\left(T_{(i)}\right)\exp(\hat{\boldsymbol{\beta}}'\mathbf{z}_{gj})\right]^{-1}$$

Here, $T_{(i)}$ are the ordered event times in group g, $Y_{gj}(T_{(i)})$ is an indicator of whether the jth individual from group g is in the risk set at time $T_{(i)}$, $\mathbf{z}_{gj}$ is the covariate vector for the jth individual from group g, and n_g is the number of subjects in group g. This leads to the following estimate of the group-specific survival function at a given value of the covariate vector $\mathbf{z}$

$$\hat{S}_g(t|\mathbf{z}) = \exp\left(-e^{\hat{\boldsymbol{\beta}}'\mathbf{z}}\hat{\Lambda}_{0g}(t)\right) \tag{4}$$

The survival estimates in equation (4) can be integrated to provide estimates of RMST and the difference in RMST between treatment groups at a given value of $\mathbf{z}$. Due to the non-linearity of the model, group differences in RMST will vary for different values of $\mathbf{z}$, *unlike* standard ANCOVA in the linear case. Karrison[21] therefore proposed averaging over the marginal covariate distributions across both treatment arms

$$\hat{S}_g\cdot(t) = \frac{1}{n_1 + n_2}\sum_{g=1}^{2}\sum_{j=1}^{n_g}\hat{S}_g\left(t|\mathbf{z}_{gj}\right)$$

to obtain an overall adjusted treatment difference

$$\hat{\delta} = \hat{\mu}_{t^*1} - \hat{\mu}_{t^*2} = \int_0^{t^*}\hat{S}_1\cdot(t)dt - \int_0^{t^*}\hat{S}_2\cdot(t)dt$$

Chen and Tsiatis[23] showed that $\hat{\delta}$ can, in fact, be interpreted as an estimate of the average causal treatment effect

$$\delta = \int_0^{t^*} E_Z[S_1(t|\mathbf{z})]dt - \int_0^{t^*} E_Z[S_2(t|\mathbf{z})]dt \tag{5}$$

The large sample variance of $\hat{\delta}$ can be obtained by the delta method and is derived in the cited papers. Of note, Karrison and Zucker conditioned on the covariates, whereas Chen and Tsiatis treated the covariates as random, which gives rise to an additional variance component. We have included this additional component in our calculations. Chen and Tsiatis also considered a more general model in which both the baseline hazard and the regression coefficients for the covariates are allowed to vary by treatment; however, our preference is for the more parsimonious model (3).

Simulation study

We conducted a simulation study to evaluate the performance of unadjusted and adjusted RMST comparisons in randomized, two-arm clinical trials. In all simulations, data were generated from a Weibull model with a pre-specified treatment effect parametrized by θ in the case of a proportional hazards treatment effect and one or more covariates. For non-proportional hazards treatment effects, different scale and shape parameters were specified for each treatment arm. Unadjusted models and models adjusting for the covariates were fitted in each data set. RMST estimates from these simulations were then used to investigate the effect of covariate adjustment and model misspecification on bias, coverage rates, power and efficiency (relative to the unadjusted estimator). Results from fitting Cox regression models were also generated. Simulation scenarios are summarized in Supplementary Table S1.

True survival times (T_i^0) were drawn with underlying survival function $S(t) = \exp(-\alpha t^{\gamma})$, where α is the Weibull scale parameter and γ is the shape parameter. We simulated RCTs with uniform accrual over 5 years followed by two additional years of follow-up; thus, censoring times were distributed uniformly, $c_i \sim Unif(2,7)$. The observed survival time was taken as $T_i = \min(T_i^0, c_i)$, with the censoring indicator denoting whether the event was observed ($\Delta_i = 1$) or censored ($\Delta_i = 0$). We investigated effects of covariate adjustment for two types of covariates: predictive covariates Z_j, where $\beta_j \neq 0$ in the true model, and unrelated covariates X_j. Data were generated so that the effect of Z_j on survival satisfied the proportional hazards assumption in all simulations, whereas the effect of treatment (I_{TRT}) satisfied the proportional hazards assumption in Scenarios 1, 4, and 5, and was non-proportional hazards in Scenarios 2 and 3. Restriction point was set at $t^* = 5$ in all simulations. Of note, even though t^* was 5 years, to maintain efficiency for the estimation of covariate effects in adjusted models, deaths and censorings occurring after 5 years were included "as is" (i.e. were not censored at 5 years).

In Scenarios 1–4, we investigated the effect of covariate adjustment when the true model has only one or two prognostic covariates. In Scenario 5, we examined the effect of covariate adjustment when the true data generating mechanism involves multiple correlated prognostic covariates with varying degrees of correlation and magnitude of the effect on survival. $R = 3000$ replications were performed for each scenario. Figure 1 shows the true survival curves for each of the five scenarios, with the covariate(s) set to their expected or representative value(s).

Scenario 1

Survival times were generated under a proportional hazards treatment effect with a single prognostic covariate Z_1 and a single non-prognostic covariate X_1. For $t^* = 5$ years, the true difference in restricted means (average causal treatment effect from equation (5)) in the non-null case is $\delta = .90$ years. Simulation results are presented in Table 1 under the null hypothesis of no treatment effect, that is, when $\theta = 0$. Table 2 presents simulation results when the treatment effect is $\theta = \ln(2)$. The total sample sizes were $N = 100$, 130, and 150 ($n = 50$, 65, and 75 per arm). Results are presented for four sets of models: unadjusted, adjusted for the prognostic covariate Z_1, adjusted for the unrelated covariate X_1, and adjusted for both Z_1 and X_1.

We found that in all models, the estimates of the treatment effect are essentially unbiased. The average model-based standard errors (ASE = mean over R replications of the estimated standard error of $\hat{\delta}$) are close to the empirical standard errors (ESE = standard deviation of $\hat{\delta}$ across R replications). Rejection rates under the null hypothesis and coverage rates under the alternative are close to the nominal 5% and 95% levels, respectively. In Table 2, adjustment for Z_1 or both Z_1 and X_1 leads to a 17%–20% improvement in efficiency relative to the unadjusted estimator, while adjustment for only the non-prognostic covariate X_1 leads to no increase in efficiency. Correspondingly, adjustment for Z_1 (or both Z_1 and X_1) increases power, whereas adjustment for the unrelated covariate X_1 results in no change in power compared to the unadjusted estimate. Of note, this is in contrast to Cox regression analysis (Supplementary Table S2), where the unadjusted estimates of the log hazard ratio, as well as estimates from models that adjust for X_1 alone, are "biased" toward zero, and the standard error of the treatment effect estimate does not decrease when adjusting for prognostic covariate Z_1. (As discussed above, "bias" and coverage rates here are with respect to the parameter θ in the model incorporating Z_1.) However, the "bias" is removed and power is increased when the correct model is fitted. Moreover, power from fitting the Cox regression model was similar to the power obtained for the

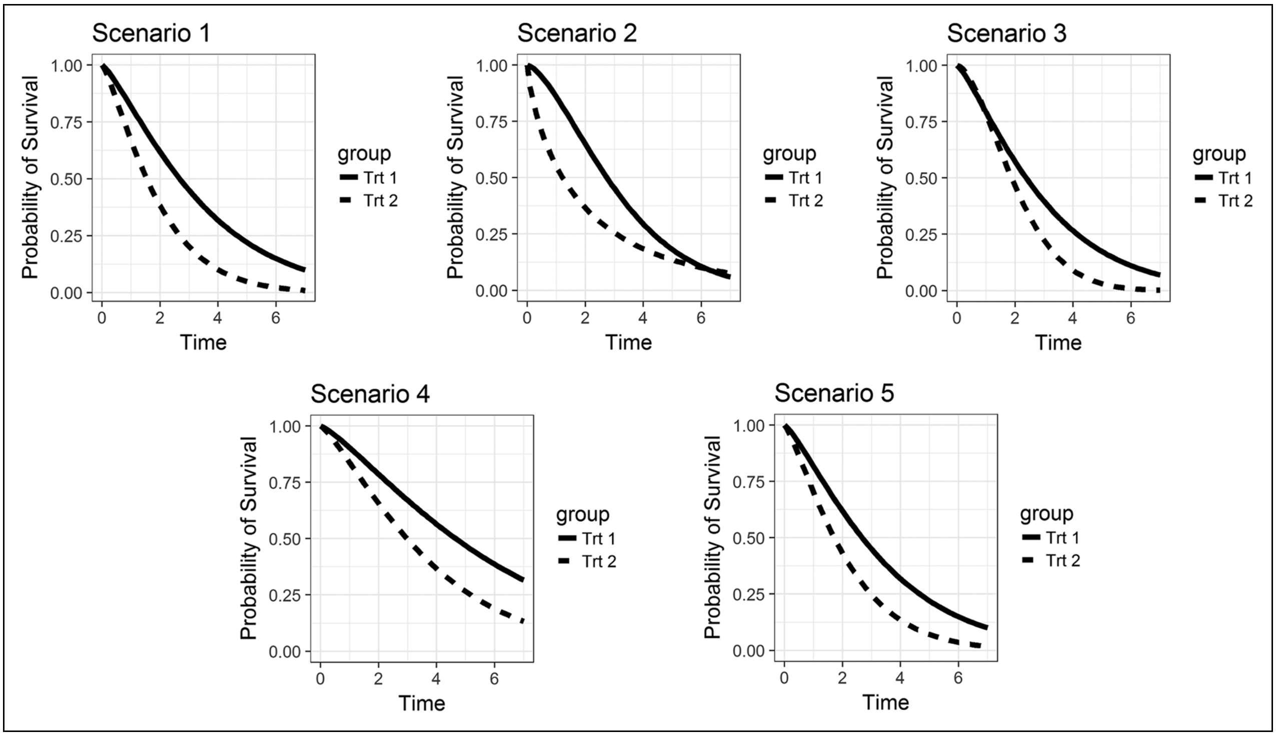


Figure 1. True survival curves at expected values of the covariates (for Scenario 4, the binary covariate Z_2 is set to 0).

Table 1. Scenario 1: null hypothesis case RMST ($\delta = 0$, Z_1 prognostic, X_1 not prognostic).

N	Model	Bias ($\hat{\delta}$)	ESE	ASE	Rej rate (%)
100	Unadj RMST	.0094	.3397	.3475	4.5
	Adj RMST				
	Z_1	.0110	.3170	.3212	5.0
	X_1	.0109	.3390	.3420	4.7
	Z_1, X_1	.0122	.3205	.3213	5.4
130	Unadj RMST	.0028	.3126	.3048	5.7
	Adj RMST				
	Z_1	.0002	.2928	.2826	5.6
	X_1	.0025	.3125	.3011	5.9
	Z_1, X_1	.0000	.2954	.2827	5.8
150	Unadj RMST	−.0071	.2822	.2830	5.1
	Adj RMST				
	Z_1	−.0071	.2610	.2629	4.6
	X_1	−.0057	.2805	.2801	5.1
	Z_1, X_1	−.0058	.2617	.2630	4.8

ESE: empirical standard error; ASE: average model-based standard error; Unadj: unadjusted; Adj: adjusted; Rej rate: rejection rate; RMST: restricted mean survival time.
Average censoring rate = 30.5%.

restricted means analysis for all estimators—both unadjusted and adjusted.

Scenario 2

Here, we generate survival times in each treatment group from Weibull distributions with different scale and shape parameters, thus allowing the treatment effect to be non-proportional hazards. Setting $\alpha_1 = 0.18, \gamma_1 = 1.50,\ \alpha_2 = 0.20, \gamma_2 = 0.75$, and $\beta_1 = \ln(2)$ produces survival curves that separate early but converge at 5 years (decreasing hazard ratio), with a true $\delta = 0.96$. Simulation results in Table 3 show that both unadjusted and adjusted estimates of δ are unbiased, with coverage rates close to their nominal 95% levels. Adjustment for the prognostic covariate Z_1

Table 2. Scenario 1: proportional hazards treatment effect, RMST (δ = .8994, Z_1prognostic, X_1 notprognostic).

N	Model	Bias ($\hat{\delta}$)	ESE	ASE	Coverage rate (%)	Power (%)	Eff
100	Unadj RMST	–.0028	.3259	.3221	94.7	78.5	–
	Adj RMST						
	Z_1	–.0126	.2970	.2974	95.0	84.8	1.20
	X_1	–.0138	.3246	.3203	94.5	78.2	1.01
	Z_1, X_1	–.0123	.2992	.2976	94.9	84.1	1.19
130	Unadj RMST	.0059	.2840	.2819	94.3	88.4	–
	Adj RMST						
	Z_1	–.0046	.2605	.2604	95.0	92.5	1.19
	X_1	–.0035	.2819	.2806	94.3	88.2	1.01
	Z_1, X_1	–.0049	.2614	.2604	94.7	92.5	1.18
150	Unadj RMST	.0057	.2702	.2627	94.0	91.9	–
	Adj RMST						
	Z_1	–.0004	.2482	.2429	94.3	95.2	1.19
	X_1	–.0014	.2696	.2617	94.3	91.9	1.00
	Z_1, X_1	.0006	.2496	.2429	94.2	95.0	1.17

ESE: empirical standard error; ASE: average model-based standard error; Unadj: unadjusted; Adj: adjusted; Eff: efficiency; RMST: restricted mean survival time.
Average censoring rate = 21.6%.

Table 3. Scenario 2: non-proportional hazards treatment effect (early difference) RMST (δ = .9628, Z_1 prognostic, X_1 not prognostic).

N	Model	Bias ($\hat{\delta}$)	ESE	ASE	Coverage rate (%)	Power (%)	Eff
100	Unadj RMST	–.0006	.3466	.3434	94.6	78.5	–
	Adj RMST						
	Z_1	–.0038	.3196	.3154	94.2	84.9	1.18
	X_1	–.0097	.3459	.3404	94.4	78.6	1.00
	Z_1, X_1	–.0050	.3226	.3156	94.1	84.6	1.15
130	Unadj RMST	–.0044	.2991	.3018	94.9	88.3	–
	Adj RMST						
	Z_1	–.0070	.2767	.2777	94.8	93.3	1.17
	X_1	–.0113	.2966	.2997	95.0	88.5	1.02
	Z_1, X_1	–.0071	.2781	.2779	94.5	93.4	1.16
150	Unadj RMST	.0031	.2785	.2802	95.5	92.9	–
	Adj RMST						
	Z_1	–.0049	.2532	.2588	95.5	96.3	1.21
	X_1	–.0027	.2770	.2785	95.6	93.1	1.01
	Z_1, X_1	–.0051	.2538	.2588	95.5	96.3	1.20

ESE: empirical standard error; ASE: average model-based standard error; Unadj: unadjusted; Adj: adjusted; Eff: efficiency; RMST: restricted mean survival time.
Average censoring rate = 23.9%.

(or both Z_1 and X_1) leads to an efficiency increase of 15%–21% relative to the unadjusted estimator, as well as increased power, whereas adjustment for X_1 alone provides no improvement in either. Supplementary Table S3 shows simulation results from fitting Cox proportional hazards models in this setting. In Cox models, adjusting for Z_1 does not reduce standard errors but does increase power. We also found that the power from the Cox regression analysis, whether unadjusted or adjusted, is lower than the corresponding power for the comparison of restricted means.

Scenario 3

In this scenario, the treatment effect is again non-proportional hazards, such that the survival curves are similar over the first year and then separate (increasing hazard ratio), with a true δ = .50 years. Sample sizes were increased to N = 150, N = 200, and N = 250 (75, 100, and 125 per arm). The simulations in Table 4 again show that all estimates are unbiased, and when comparing restricted means between groups, both efficiency and power are increased if the model includes

Table 4. Scenario 3: non-proportional hazards treatment effect (late difference), RMST (δ = .4972, Z_1 prognostic, X_1 not prognostic).

N	Model	Bias ($\hat{\delta}$)	ESE	ASE	Coverage rate (%)	Power (%)	Eff
150	Unadj RMST	.0106	.2506	.2502	94.5	52.6	–
	Adj RMST						
	Z_1	.0041	.2278	.2313	95.1	58.2	1.21
	X_1	.0048	.2495	.2497	94.7	51.3	1.01
	Z_1, X_1	.0041	.2288	.2314	95.1	58.3	1.20
200	Unadj RMST	.0054	.2128	.2168	95.5	63.7	–
	Adj RMST						
	Z_1	.0038	.1963	.2002	95.5	70.3	1.18
	X_1	.0008	.2125	.2165	95.4	63.1	1.00
	Z_1, X_1	.0033	.1975	.2003	95.3	70.1	1.16
250	Unadj RMST	.0033	.1931	.1940	94.5	73.0	–
	Adj RMST						
	Z_1	.0005	.1793	.1791	95.1	78.7	1.16
	X_1	.0005	.1928	.1937	94.5	72.5	1.00
	Z_1, X_1	.0011	.1801	.1791	94.8	78.9	1.15

ESE: empirical standard error; ASE: average model-based standard error; Unadj: unadjusted; Adj: adjusted; Eff: efficiency; RMST: restricted mean survival time.
Average censoring rate = 19.4%.

Table 5. Scenario 4: proportional hazards treatment effect RMST (δ = .7007, Z_1, Z_2 prognostic, X_1 not prognostic).

N	Model	Bias ($\hat{\delta}$)	ESE	ASE	Coverage rate (%)	Power (%)	Eff
200	Unadj RMST	−.0030	.2401	.2431	95.2	81.6	–
	Adj RMST						
	Z_1	−.0088	.2334	.2343	95.0	83.4	1.06
	X_1	−.0072	.2389	.2407	95.1	81.9	1.01
	Z_1, X_1	−.0088	.2338	.2342	95.0	83.4	1.05
	Z_2	−.0062	.2364	.2374	95.3	82.8	1.03
	Z_2, X_1	−.0062	.2367	.2374	95.1	82.6	1.03
	Z_1, Z_2	−.0076	.2310	.2307	94.9	84.3	1.08
	Z_1, Z_2, X_1	−.0075	.2314	.2307	95.0	84.5	1.08

ESE: empirical standard error; ASE: average model-based standard error; Unadj: unadjusted; Adj: adjusted; Eff: efficiency; RMST: restricted mean survival time.
Average censoring rate = 37.0%.

the prognostic covariate Z_1. In Cox regression models (Supplementary Table S4), standard errors are not reduced but power increases when Z_1 is included in the model. In this case, Cox regression analysis has *higher* power than comparison of restricted means. Our findings about power for RMST estimation relative to Cox regression under proportional hazards, early and late treatment difference alternatives mirror the findings reported in Karrison.[15]

Scenario 4

We return to considering a proportional hazards treatment effect but with two prognostic covariates and one non-prognostic covariate. We generated $Z_1 \sim N(0, 1)$ and Z_2 independently as a binary covariate taking values 0 and 1 with probability .5. $X_1 \sim Unif(0, 2)$ was generated independent of Z_1 and Z_2, and we set $\theta = \ln(1.75)$, $\beta_1 = \ln(1.33)$, and $\beta_2 = \ln(1.5)$. With these parameter values, the true average causal treatment effect at $t^* = 5$ years is $\delta = .70$ years. The sample size was $N = 200$ ($n = 100$ per treatment group). Table 5 shows efficiency and power, along with the other metrics, for the unadjusted model, as well as models that adjust for different combinations of covariates (e.g. only Z_1, only X_1, Z_1 and X_1, etc.). Compared to the unadjusted estimator, adjusting for X_1 does not increase efficiency. Adjusting for Z_1 increases efficiency by about 6%, adjusting for Z_2 increases efficiency by about 3%, and adjusting for both prognostic factors increases efficiency by 8%. Correspondingly, the power is increased slightly from 82% for the unadjusted estimator to a little over 84% when adjusting for both Z_1 and Z_2. This simulation suggests that efficiency gains may be relatively minor when covariates have only modest prognostic effects. Supplementary Table S5

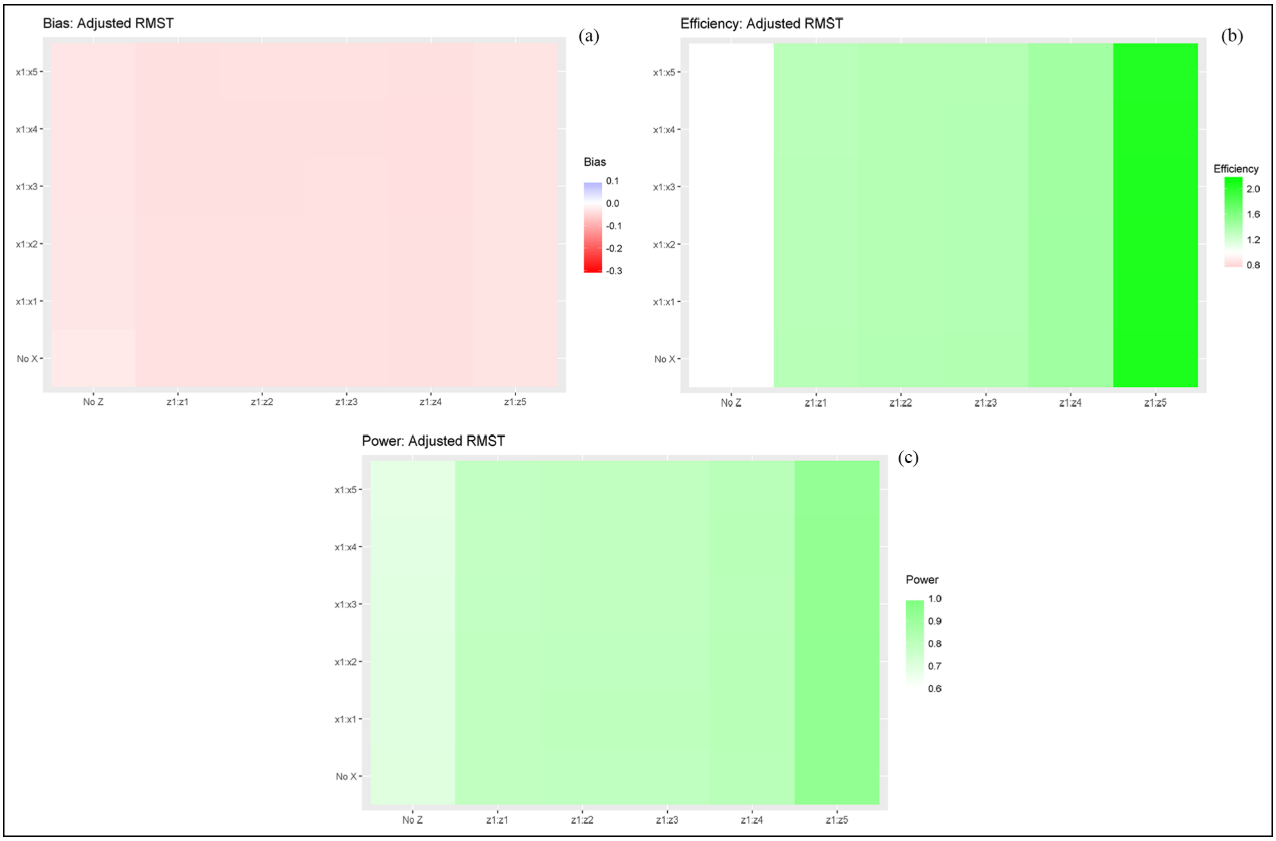


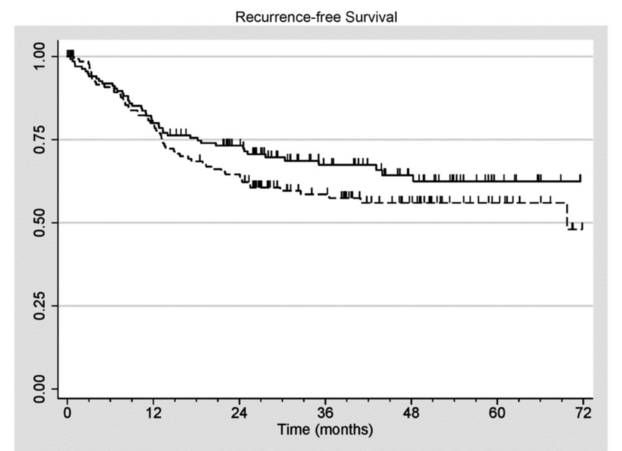
Figure 2. Scenario 5 RMST: ($\delta = .5973, Z_1, \ldots, Z_5$ prognostic, $X_1, \ldots, X_5$ not prognostic). Average censoring rate = 27.2%, (a) Bias; (b) Efficiency; (c) Power

shows that in the case of Cox regression, "bias" is reduced and power increased when the prognostic covariates are added, but again the magnitudes of the gains are modest due to the smaller covariate effects.

Scenario 5

In the last scenario, we consider five correlated prognostic covariates ($Z_1, \ldots, Z_5$) and five unrelated covariates ($X_1, \ldots, X_5$) that are not associated with outcome, and a treatment effect that satisfies the proportional hazards assumption. Both sets of covariates were generated from a multivariate normal distribution with mean and correlation structure as shown in Table S1. Z_1 is not correlated with $Z_2, \ldots, Z_5$, but $Z_2, \ldots, Z_5$ are correlated with each other, with ρ ranging from .1 to .3, and similarly for $X_1, \ldots, X_5$. Survival times were generated from a Weibull distribution with $\alpha = .16$ and $\gamma = 1.25$, and the magnitude of the covariate effects decreased from Z_1 to Z_5. The treatment effect was set at $\theta = \ln(1.75)$. The sample size was $N = 200$ ($n = 100$ per treatment group). The true average causal treatment effect at $t^* = 5$ years is $\delta = .60$. Figure 2(a)–(c) are heat maps showing bias, efficiency, and power that result from models adjusting for various combinations of covariates. The lower left-hand corner of each map corresponds to the unadjusted model ("No Z," "No X"). Moving up vertically are models adjusting for $X_1, X_1 + X_2, \ldots, X_1 + \cdots + X_5$, whereas moving horizontally to the right adds $Z_1, Z_1 + Z_2, \ldots, Z_1 + \cdots + Z_5$. The upper right-hand corner includes all 10 covariates. All models yield estimates with little or no bias, while progressively adjusting for the prognostic covariates increases statistical efficiency and power. Heat maps from fitting Cox regression models under this scenario are shown in Supplementary Figures S1A and S1B. We again see that model misspecification in the Cox regression analysis leads to negatively "biased" estimates. Power is noticeably increased as prognostic covariates are added to the model because the negative "bias" is gradually removed.

The covariate effects in Scenario 5 are somewhat large. For example, $\beta_1 = \ln(2)$ implies a hazard ratio of 4 for Z_1 equal to one standard deviation below the mean compared to one standard deviation above the mean. As a result, the efficiency for the RMST comparison relative to the unadjusted estimate ranged from 1.34 with the inclusion of Z_1 to 2.07 with the inclusion of all five covariates. We re-ran the simulations for Scenario 5 with smaller covariate effects, that is, we set the coefficients to ln(1.5), ln(1.25), ln(1.1), ln(.9), and ln(.75) for $\beta_1 - \beta_5$, respectively. Not unexpectedly, the efficiency gains were more modest, ranging from 1.16



I+CRT							
At risk:	138	108	86	54	35	14	4
Events:	27	9	6	2	1	0	
Cens:	3	13	26	17	20	10	
CRT							
At risk:	135	103	83	53	32	14	3
Events:	27	19	7	2	0	1	
Cens:	5	1	23	19	18	10	

Figure 3. Recurrence-free survival in DeCIDE trial of induction therapy plus chemoradiotherapy (I + CRT) versus CRT alone in patients with head-and-neck cancer. Solid: I + CRT (n_1 = 138), Dashed: CRT (n_2 = 135).

with the inclusion of Z_1 to 1.29 when all five prognostic covariates were included in the model. The heat map for efficiency is given in Supplementary Figure S2.

Example

As an example, we analyze data from the DeCIDE trial,[24] an RCT of induction therapy plus chemoradiotherapy (I + CRT) versus chemoradiotherapy (CRT) alone in patients with locally advanced squamous cell carcinoma of the head and neck. Patients with non-metastatic N2 or N3 disease were randomized to receive either two cycles of induction therapy followed by five cycles of chemoradiotherapy or five cycles of chemoradiotherapy only. A total of 273 evaluable patients were enrolled and followed for up to 7 years. The Kaplan–Meier curves for recurrence-free survival, defined as the time from randomization until disease recurrence or death from any cause, are shown in Figure 3. Recurrence-free survival is higher in the induction therapy plus chemoradiotherapy arm after 1 year, but the difference is not statistically significant by the log-rank test (p = .16). Estimating RMST at t^* = 5 years gives the following unadjusted RMST estimates (±standard error (SE)):

Unadjusted

$$\hat{\mu}_{I\,+\,CRT} = 3.645 \pm .169 \text{ years}, \ 3.645/5 = 73\%$$

$$\hat{\mu}_{CRT} = 3.322 \pm .177 \text{ years}, \ 3.322/5 = 66\%$$

$$\hat{\delta} = .323 \pm .245 \text{ years}, \ \hat{\mu}_{I\,+\,CRT}/\hat{\mu}_{CRT} = 1.10, \ p = .19$$

Thus, in the induction therapy plus chemoradiotherapy arm, RMST restricted to 5 years was estimated to be 3.64 years, and patients achieved 73% of potential recurrence-free life years (over a 5-year horizon) compared to 3.32 years and 66% in the chemoradiotherapy-only arm. The absolute difference in restricted means is .32 years (ratio 1.10), but is not statistically significant (p = .19). The p-value is very close to the p-value from the log-rank test.

Next, we obtain the RMST estimate adjusting for five prognostic covariates that were all significantly associated with recurrence-free survival in univariate analyses, that is, Karnofsky performance score, T-stage, N-stage, age, and smoking status.

Adjusted

$$\hat{\mu}_{I\,+\,CRT} = 3.601 \pm .156 \text{ years}, \ 3.601/5 = 72\%$$

$$\hat{\mu}_{CRT} = 3.352 \pm .158 \text{ years}, \ 3.352/5 = 67\%$$

$$\hat{\delta} = .249 \pm .223 \text{ years}, \ \hat{\mu}_{I\,+\,CRT}/\hat{\mu}_{CRT} = 1.07, \ p = .26$$

Here, despite the increase in precision, adjustment has reduced the estimated difference in restricted means and increased the p-value. This is because the induction therapy plus chemoradiotherapy arm was slightly favored on these covariates. Of interest, similar conclusions are obtained from fitting a Cox proportional hazards regression model to these data: unadjusted $\hat{\beta} = -.278 \pm .200$, hazard ratio (I + CRT/CRT) = .76, p = .16; adjusted $\hat{\beta} = -.232 \pm .204$, hazard ratio = .79, p = .25.

Discussion

Our simulation study suggests that analysis of restricted means based on the stratified Cox model (3) is similar to ANCOVA for linear models, in that adjusting for covariates associated with the outcome provides increased precision for the treatment effect contrast, whereas adjustment for non-prognostic covariates produces no improvement. Our findings suggest that incorporating covariates into the model can improve precision if they are appropriately chosen. A conservative approach to design clinical trials that compare RMST could be to power the study based on the expected precision of the unadjusted estimator, and then to incorporate covariates into the final analysis to narrow the confidence interval width and increase power. However, there can be downsides to this strategy. As shown by Beach and Meier,[25] adjustment for covariates in RCTs affords the analyst the opportunity to select the model that provides the strongest evidence for a treatment effect—so-called "p-value shopping." One solution to this problem is to pre-specify in the protocol the set of covariates that one will include in the model based on a priori knowledge about which factors are likely to affect survival, that is, known prognostic factors. Alternatively, Tsiatis et al.[26] have developed a strategy for covariate adjustment that avoids this pitfall and which could potentially be adapted to RMST.

A nice feature of the analysis of restricted means, as suggested by our simulation studies, is that unadjusted estimates, as well as estimates from models that include only some of the true prognostic factors, show little or no bias. This implies that while some efficiency may be lost, the treatment effect estimator is centered at the same target even when the model is misspecified and influential covariates are omitted. In addition, the estimand can be interpreted as the average causal treatment effect, and its interpretation does not rely on proportional hazards or other parametric assumptions.

Finally, we reemphasize that RMST estimates require careful interpretation. If the survival estimates are at or near zero toward the end of the follow-up period, the restricted mean will be close in magnitude to the overall mean. But this is frequently not the case in clinical trials where the follow-up time can be relatively limited, resulting in high censoring rates, and such that survival estimates remain above 25% or even above the 50th percentile as, for example, in the DeCIDE trial. If survival rates differ at the end of the follow-up period and the true curves remain separated, the difference in restricted means will underestimate—potentially seriously underestimate—the difference in overall means. Thus, RMST informs us only about the survival experience up to the limit of observation. What else could it do? Only parametric assumptions or extrapolation beyond the observation period would give us estimates of the overall mean, and few would likely want to rely on such an approach. Nonetheless, with these caveats in mind, analysis of RMST can provide informative results about the effects of treatment on survival in clinical trials and be a useful complement to standard methods.

Acknowledgements

The authors thank David Zucker for use of his SAS macros and the reviewers for their helpful comments.

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

This work was supported by National Cancer Institute (Grant/Award Number: "P30CA014599" and "P30CA060553"). The content is solely the responsibility of the authors and does not necessarily represent the views of the National Institutes of Health.

References

1. Cox DR. Regression models and life tables (with discussion). *J Roy Stat Soc B Met* 1972; 34: 187–220.
2. Grambsch PM and Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; 81: 515–526.
3. Kalbfleisch JD and Prentice RL. *The statistical analysis of failure time data*. New York: John Wiley & Sons, 1980.
4. Vansteelandt S, Martinussen T and Tchetgen EJT. On adjustment for auxiliary covariates in additive hazard models for the analysis of randomized experiments. *Biometrika* 2014; 101: 237–244.
5. Buckley J and James I. Linear regression with censored data. *Biometrika* 1979; 66: 429–436.
6. Chiou SH, Kang S and Yan J. Fitting accelerated failure time models in routine survival analysis with R package aftgee. *J Stat Softw* 2014; 61: 1–23.
7. Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *J Hyg* 1949; 47: 188–189.
8. Kaplan EL and Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958; 53: 457–481.
9. Meier P. Estimation of a distribution function from incomplete observations. In: Gani J (ed.) *Perspectives in probability and statistics: papers in honour of M.S. Bartlett*. New York: Academic Press, 1975, pp. 67–87.
10. Royston P and Parmar MKB. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 2013; 13: 152.
11. Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *J Clin Oncol* 2014; 32: 2380–2385.
12. Trinquart L, Jacot J, Conner SC, et al. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *J Clin Oncol* 2016; 34: 1813–1819.
13. Chappell R and Zhu X. Describing differences in survival curves. *JAMA Oncol* 2016; 2: 906–907.
14. Uno H, Wittes J, Fu H, et al. Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies. *Ann Intern Med* 2015; 163: 127–134.
15. 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: 151–167.
16. Robinson LD and Jewell NP. Some surprising results about covariate adjustment in logistic regression models. *Int Stat Rev* 1991; 59: 227–240.
17. Ford I, Norrie J and Ahmadi S. Model inconsistency, illustrated by the Cox proportional hazards model. *Stat Med* 1995; 14: 735–746.
18. 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.

19. Hauck WW, Anderson S and Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? *Control Clin Trials* 1998; 19: 249–256.
20. Schoenfeld DA and Borenstein M. Calculating the power or sample size for the logistic and proportional hazards models. *J Stat Comput Sim* 2005; 75: 771–785.
21. Karrison T. Restricted mean life with adjustment for covariates. *J Am Stat Assoc* 1987; 82: 1169–1176.
22. Zucker DM. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *J Am Stat Assoc* 1998; 93: 702–709.
23. Chen P-Y and Tsiatis AA. Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics* 2001; 57: 1030–1038.
24. Cohen EE, Karrison TG, Kocherginsky M, et al. Phase III randomized trial of induction chemotherapy in patients with N2 or N3 locally advanced head and neck cancer. *J Clin Oncol* 2014; 32: 2735–2743.
25. Beach ML and Meier P. Choosing covariates in the analysis of clinical trials. *Control Clin Trials* 1989; 10: 161S–175S.
26. 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.