The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
15, Number 4, pp. 1098–1117

Estimating the treatment effect in a clinical trial using difference in restricted mean survival time

Patrick Royston
MRC Clinical Trials Unit
University College London
London, UK
j.royston@ucl.ac.uk

Abstract. The causal effect of a new medical treatment compared with a standard regimen is best assessed in a randomized controlled trial setting. When the main outcome is time to some event of interest, such as death, studies often use the hazard ratio to describe the treatment effect. Typically, proportional hazards are assumed. Here I discuss several significant disadvantages of using the hazard ratio, including its vulnerability to the proportionality assumption, its relative nature, and its lack of relationship with time-to-event or survival probabilities. I describe the use of restricted mean survival time as an alternative outcome measure in time-to-event trials. With this method, the treatment effect is defined as the difference in restricted mean between the trial arms. I suggest the use of Royston and Parmar's (2002, *Statistics in Medicine* 21: 2175–2197) class of flexible parametric models, implemented through the command `stpm2` (Lambert and Royston [2009, *Stata Journal* 9: 265–290] and Andersson and Lambert [2012, *Stata Journal* 12: 623–638]), to estimate the required quantities. With this approach, proportional hazards are not assumed. I describe a new command, `strmst`, for implementing these calculations. This method supports "direct" adjustment for covariates by using marginalization over their observed distribution, and it supports estimation of treatment effects conditional on fixed values of covariates. I illustrate the methodology using data from a trial in primary biliary cirrhosis. I provide an example that demonstrates the importance of understanding the relationship between the treatment effect, the prognosis of the disease outcome, and the often-neglected time domain.

Keywords: st0415, strmst, randomized trial, time-to-event data, treatment effect, restricted mean, flexible parametric model, nonproportional hazards

1 Introduction

In medicine, the causal effect of a new research treatment compared with that of an existing standard treatment is classically assessed using selected patients in a randomized controlled trial. When the main outcome is time to some event of interest, such as death, the measure usually used to describe the treatment effect is the hazard ratio (HR). In many trials, past and present, the HR appears reasonably constant during the follow-up period of the trial, which suggests that it is a useful, time-invariant summary of a treatment effect.

However, the HR has at least two significant drawbacks (Royston and Parmar 2013): i) it is a relative measure, so it tells us nothing about the absolute improvement in survival probability or time to event due to the research treatment, and ii) it is supposedly independent of time, so it effectively suppresses the important time dimension of a trial as a possible factor in a treatment's success or failure. Regarding i), a relative measure is advantageous when one wants to summarize evidence for a treatment effect over several studies, but it is not easy to interpret in practical terms because it provides no information about the effectiveness of a treatment for survival time or survival probability. I give an example of the latter in the next paragraph. Regarding ii), the treatment effects in a large trial that lasts one year and in a small trial that lasts much longer may differ in clinical importance, even when the HRs and numbers of events are the same.

When the survival probability scale is considered, the treatment effect (represented by a constant HR) for a group whose survival is very high or very low after a given time interval may be quite different from the treatment effect for a group whose survival is "in the middle". For example, following surgery to remove the tumor in primary breast cancer, axillary node-negative patients may have a five-year overall survival of around 90%. Giving chemotherapy with an HR of 0.75 to such patients would increase the five-year overall survival to $100\left(0.90^{0.75}\right) = 92.4\%$, which is a rather small increment despite the impressive 25% reduction in the mortality rate (hazard function). By contrast, the five-year overall survival for poor-prognosis patients with more than 9 positive lymph nodes may be about 55%. Chemotherapy with HR $= 0.75$ would improve this to 64%, which seems to indicate a much more worthwhile return for the toxicity known to be associated with the treatment.

I have illustrated the effect of an HR on survival probabilities at a particular time point. However, survival probabilities are not necessarily easy for patients and physicians to comprehend and compare. Some type of average time to event may be easier. One option, on which I focus exclusively here, is restricted mean survival time (RMST).

The RMST concept is not new (Irwin 1949) and has occasionally been used in clinical trials (for example, Yusuf et al. [1994]). However, following the recent seminal work of Per Kragh Andersen and others (Andersen, Hansen, and Klein 2004; Andersen and Pohar Perme 2010; Parner and Andersen 2010; Overgaard, Andersen, and Parner 2015), RMST has gained new impetus as an alternative approach to the HR as a summary measure of a treatment effect (Royston and Parmar 2011, 2013). On a related point, it cannot be overstated how critical it is to prespecify in the trial protocol the statistical analysis that will be performed on the data, once mature. This is essentially a regulatory requirement in the research governance of clinical trials. Hence, in this article, I stress the prespecification of a flexible parametric model (FPM) with $(3, 1)$ degrees of freedom (d.f.) (possibly accompanied by a sensitivity analysis) as a reasonably robust route to estimating RMST and its difference. The FPM method is powerful also because it supports estimation of effects directly adjusted for prognostic and stratification factors, which is a common requirement in the primary analysis of clinical trials data.

To understand RMST, let's suppose we have data from a trial in which all patients were followed up until death. (Such a trial may not exist!) We might be interested in comparing the mean survival time of the randomized groups. We could easily do this using standard methods, for example, by calculating the difference in mean survival [together with its 95% confidence interval (CI)] and performing a two-sample t test or Mann–Whitney test. If we were interested in the five-year survival, a natural approach would be to truncate survival times beyond five years at five years. We could then perform the same analytic procedures (although we would probably prefer the Mann–Whitney test to the t test). The resulting means in each group would now be "restricted means" because they would apply to only the interval $[0, 5]$ year. A patient-centered interpretation of RMST at five years could be as follows: "If I just focus on the next five years, I can expect to live about four more years, and if I take the new treatment, I may live about six months longer than if I opt for the standard treatment".

In a real trial, we have censoring of some times to event. A consequence of this is that we can no longer correctly compute and compare arithmetic mean survival times. However, despite the censoring, we can still calculate the RMST. We can do this "non-parametrically", for example, by using Kaplan–Meier survival curves or via "pseudo-observations" (Parner and Andersen 2010). However, for present purposes, we prefer to work with FPMs, also known as Royston–Parmar models (see Royston and Lambert [2011]). FPMs provide a unified approach to model fitting and estimation of treatment effects in a flexible framework. In the next section, I briefly describe how RMST is calculated with an FPM.

RMST for a given population critically depends on two things: i) the time point (t^*) chosen for evaluation and ii) the shape of the survival curve over time. Difference in RMST between treatment groups additionally depends on the magnitude of the HR, whether it be constant or time dependent. It is important to choose a clinically relevant value of t^* that typically reflects the outcome over a key period of exposure to the disease and treatments, such as five-year survival for some cancers. Because we wish to avoid extrapolation, t^* must be, at the very most, no greater than the largest uncensored event time in the dataset.

In real trials, it is commonplace to adjust a treatment effect for design features such as stratification variables and known prognostic factors such as age. This is easily done when estimating an HR within a Cox model; the features of interest are simply included as predictors in the model. In this article, I use the method of directly adjusted survival curves (Royston and Lambert 2011, sec. 9.3) to obtain adjusted RMST values.

I describe a new command, `strmst`, that estimates the treatment effect according to RMST and its difference between trial arms. `strmst` supplies CIs and can adjust for covariates if required. The command routinely incorporates nonproportional hazards of treatment effects. Note that `strmst` requires `stpm2` (which fits FPMs) (Lambert and Royston 2009; Andersson and Lambert 2012).

The article is structured as follows. In section 2, I discuss RMST and its difference as an estimate of the treatment effect. In section 3, I describe a clinical trial in liver disease that we use as an example. In section 4, I demonstrate flexible parametric modeling of

the data and how to use `strmst` to obtain unadjusted and adjusted estimates of RMST. I show how to plot corresponding unadjusted and adjusted survival curves. Finally, I suggest a sensitivity analysis that is intended to accommodate model uncertainty in the selected FPM. In section 5, I describe the syntax and details of `strmst`. In section 6, I conclude with some brief remarks.

2 RMST and its difference

The RMST $\mu(t^*)$ at some time horizon $t^* > 0$ for a time-to-event random variable T is the expectation of the truncated survival time $X = \min(T, t^*)$. It may be shown that $\mu(t^*)$ is the area under the survival curve $S(t)$ from $t = 0$ to $t = t^*$ (Andersen and Pohar Perme 2010, Irwin 1949); that is,

$$\mu(t^*) = E(X) = E\{\min(T, t^*)\} = \int_0^{t^*} S(t)\, dt$$

When T is years to death, we may think of $\mu(t^*)$ as the t^*-year life expectancy. In a two-arm clinical trial with survival functions $S_0(t)$ and $S_1(t)$ in the control and research arms, respectively, the difference in RMST between arms is given by

$$\begin{aligned}\Delta(t^*) &= \int_0^{t^*} S_1(t)\, dt - \int_0^{t^*} S_0(t)\, dt \\ &= \int_0^{t^*} \{S_1(t) - S_0(t)\}\, dt\end{aligned}$$

That is, $\Delta(t^*)$ is the area between the survival curves. Depending on the pattern of the survival curves, $\Delta(t^*)$ may be negative, zero, or positive. An effective research treatment will have significant, positive $\Delta(t^*)$ for values of t^* of clinical interest.

Section 4 describes the use of `strmst` to estimate RMST and $\Delta(t^*)$ according to an FPM.

With $X = \min(T, t^*)$ defined as above, we provide the following derivation of its mean, which is the RMST at t^*.

Let $P()$ denote the probability, $F(t)$ the distribution function of T, $f(t)$ the density of T, and $S(t) = 1 - F(t)$ the survival function. The distribution of $\min(T, t^*)$ is $F(t)$ on the interval $[0, t^*]$ and is the point mass $P(T > t^*)$ on t^*. The integral giving the mean, namely,

$$E(X) = E\{\min(T, t^*)\} = \int \min(T, t^*)\, dF\{\min(T, t^*)\}$$

is split into the interval $[0, t^*]$, where $\min(T, t^*)$ is T with distribution $F(t)$, and the point (t^*), where $\min(T, t^*)$ is t^* with distribution $P(T > t^*)$. Thus

$$\begin{aligned} E\{\min(T, t^*)\} &= \int_0^{t^*} tf(t)\,dt + t^* P(T > t^*) \\ &= F(t^*)\,t^* - \int_0^{t^*} 1F(t)\,dt + t^* P(T > t^*) \\ &= t^* - \int_0^{t^*} F(t)\,dt = \int_0^{t^*} \{1 - F(t)\}\,dt \\ &= \int_0^{t^*} S(t)\,dt \end{aligned}$$

3 Data

Primary biliary cirrhosis (PBC) is a serious liver disease that usually results in liver failure and death. The effect of the drug azothiaprine on the survival of patients with PBC was compared with placebo in a multinational, double-blind, randomized clinical trial (Christensen et al. 1985). Between 1971 and 1977, 248 patients were randomized to receive either azothiaprine or placebo, with follow-up until 1983. After 41 (17%) cases with missing values or no patient follow-up were removed, data on 207 patients (105 deaths) were available for analysis. Relevant prognostic factors were age, log bilirubin, and albumin. As Royston, Altman, and Sauerbrei (2006) demonstrated, because of chance imbalance in the most important prognostic factor (bilirubin) between the arms of this relatively small trial, adjustment for prognostic factors had an unusually large influence on estimated treatment effects.

4 Analysis of the example dataset

4.1 Conventional analysis

Figure 1 shows Kaplan–Meier curves for the research and control arms of the PBC trial.

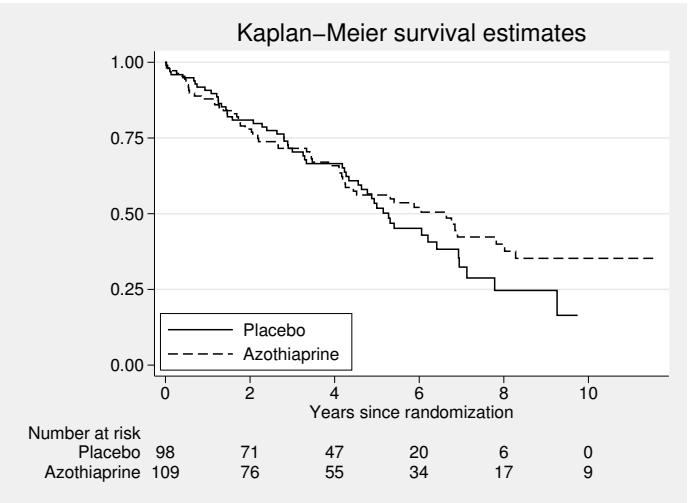

Figure 1. Kaplan–Meier curves for overall survival in the PBC trial

There is no clear difference between the survival curves. The unadjusted HR [with 95% CI] is 0.83 $[0.57, 1.22]$, $P = 0.35$. There is minor evidence of nonproportional hazards, with $P = 0.086$ according to the Grambsch–Therneau test, as performed by `estat phtest, rank`. The graph suggests that there might be a "late" treatment difference starting about five years after randomization.

A suitable prognostic model for this dataset includes the variables age, log bilirubin, and albumin. After one adjusts for the prognostic factors, the HR (CI) is 0.61 $[0.41, 0.91]$, $P = 0.016$. No significant nonproportionality of the hazards is seen for treatment ($P = 0.3$) or for the prognostic factors ($P > 0.4$). The adjusted result totally changes the interpretation of the trial outcome. The change is due to a fairly small, nonsignificant but influential difference in the highly prognostic variable log bilirubin between the arms.

4.2 Flexible parametric modeling

Royston and Parmar (2013) suggested that for the analysis of clinical trial data, a good strategy is to fit a prespecified FPM and use the model to estimate the trial-related quantities of interest. There are three major reasons: i) flexibility in prediction of relevant quantities, including adjusted survival curves; ii) transparency, to avoid data-manipulation bias ("data dredging"); and iii) fit, because the FPM with a possibly

time-dependent treatment effect will fit the data better than standard survival models. It is usually necessary to decide and document the analysis approach in the clinical protocol before obtaining and seeing the data. In my experience, a sensible FPM to use in this situation is a hazards-scaled model with 3 d.f. for the spline function representing the baseline log cumulative-hazard function and 1 d.f. for a possible time-dependent treatment effect. This is called a "$(3, 1)$ d.f." model. The 3 d.f. for the baseline specifies two knots (polynomial join points) for the restricted cubic spline function, providing considerable flexibility in the shape of the function. The 1 d.f. for time dependency means that the regression coefficient for treatment is constrained to be a linear function of log follow-up time. This simple specification allows the estimated treatment effect to increase, remain constant, or diminish over time.

I now illustrate the use of `strmst` to obtain unadjusted and adjusted RMST values and their between-arm difference in the PBC trial. First, a suitable t^* must be selected. Generally, we want to cover most of the follow-up period; otherwise, there is little point in obtaining long-term follow-up data. Here $t^* = 10$ years is probably too large because no patient in the placebo arm is still at risk by 10 years (see figure 1). We take $t^* = 8$ years. The analysis with `strmst` is straightforward.

```
. use pbc
(Primary biliary cirrhosis)
. generate ln_bilirubin = ln(bilir)
. strmst trt, tstar(8)
Log likelihood = -252.89785                     Number of obs     =        207
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
trt	-.056214	.2269794	-0.25	0.804	-.5010854	.3886574
_rcs1	1.254529	.1700067	7.38	0.000	.9213225	1.587736
_rcs2	-.1016293	.0795416	-1.28	0.201	-.257528	.0542693
_rcs3	-.0221963	.0445647	-0.50	0.618	-.1095415	.0651489
_rcs_trt1	-.2860466	.2013678	-1.42	0.155	-.6807202	.108627
_cons	-1.047456	.1629987	-6.43	0.000	-1.366927	-.7279842

The above table shows the result of fitting the $(3, 1)$ d.f. model with `strmst`. The linear predictor (`xb`) has terms for the treatment effect (`trt`) and its interaction with time (`_rcs_trt1`) and terms (`_rcs1`, `_rcs2`, `_rcs3`) corresponding to the three spline basis functions, the first of which is linear in $\ln(\_t)$, that are required for `df(3)`. The nonsignificance ($P = 0.155$) of the treatment-time interaction term suggests that there are no substantial nonproportional hazards. Nevertheless, we persist with the prespecified FPM and refrain from assessing the proportional-hazards (PH) assumption formally.

The output continues as follows:

```
Using delta method for SE and CI:
```

	Observed Est.	Delta-meth. Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
rmst1	4.9587855	.29663131	16.72	0.000	4.377399	5.5401721
rmst2	5.2990036	.28995311	18.28	0.000	4.7307057	5.8673015
dif21	.34021807	.41358548	0.82	0.411	-.47039461	1.1508307

`strmst` presents a table of the RMST estimates for placebo and azothiaprine (`rmst1` and `rmst2`, respectively) and their difference (`dif21`, azothiaprine minus placebo). The delta method (Stata's `predictnl` command) is used to obtain standard errors (SEs) via numerical derivatives. We see that the RMST estimates for placebo and azothiaprine are 5.0 and 5.3 years, respectively; this means that the average survival time over the interval $(0, 8)$ years is about 5 years. The RMST difference (`dif21`) is $\Delta\,(8) = 0.34$ (95% CI, $[-0.47, 1.15]$) years. The p-value of 0.41 comparing `dif21` with 0 is not significant at the 5% level.

We now observe the analysis with direct adjustment for the three important prognostic factors described above. Direct adjustment is done by "marginalization over the distribution of covariates" (see http://en.wikipedia.org/wiki/Marginal_distribution for an explanation). Here marginalization entails averaging predicted survival curves across all covariate patterns in the estimation sample and integrating the resulting average survival curves over $(0, t^*)$ to obtain RMST values. The required `strmst` command is simple.

```
. strmst trt, tstar(8) adjust(ln_bilirubin age albumin)
Log likelihood = -191.27809                     Number of obs     =        207
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
trt	-.4489923	.2186604	-2.05	0.040	-.8775587	-.0204258
ln_bilirubin	1.168652	.1225971	9.53	0.000	.9283657	1.408937
age	.0379697	.0120093	3.16	0.002	.014432	.0615074
albumin	-.0421403	.0180153	-2.34	0.019	-.0774497	-.0068309
_rcs1	1.479009	.1743993	8.48	0.000	1.137193	1.820825
_rcs2	-.2682558	.0883365	-3.04	0.002	-.4413921	-.0951194
_rcs3	-.109121	.0582466	-1.87	0.061	-.2232822	.0050402
_rcs_trt1	-.2104133	.2163135	-0.97	0.331	-.6343799	.2135533
_cons	-5.96387	1.24408	-4.79	0.000	-8.402222	-3.525517

```
Using delta method for SE and CI:
```

	Observed Est.	Delta-meth. Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
rmst1	4.7989807	.20651278	23.24	0.000	4.3942232	5.2037382
rmst2	5.4985933	.19608389	28.04	0.000	5.1142759	5.8829107
dif21	.69961262	.26784077	2.61	0.009	.17465433	1.2245709

We see that the effects of the three prognostic factors are significant ($P < 0.02$), particularly that for log bilirubin. The RMST values at $t^* = 8$ years are now 4.8 and 5.5 years, with $\Delta(8) = 0.70$ $(0.17, 1.22)$ years. The treatment effect has more than doubled compared with the unadjusted analysis. Furthermore, the p-value for testing $\Delta(8) = 0$ is smaller than that for testing $\ln(\text{HR}) = 0$ in the corresponding adjusted Cox model (0.0091 versus 0.016). The result suggests that the FPM analysis may have more power here.

We conclude that after we consider important prognostic factors, there is an important difference in RMST at eight years favoring azothiaprine treatment. Patients appeared to survive on average 0.7 years longer over the first 8 years with azothiaprine treatment than with placebo.

4.3 Conditional estimation

In the introduction, I alluded to the fact that the magnitude of a treatment effect on the survival scale depends on prognostic information. For example, after several years' follow-up, patients with a good breast cancer prognosis may benefit less from chemotherapy than those with a poor prognosis. The same may apply to RMST difference, as shown in the following example for the PBC trial.

Figure 2 shows the RMST difference $\Delta(8)$ as a function of centiles of the observed distribution of log bilirubin, adjusted for age and albumin.

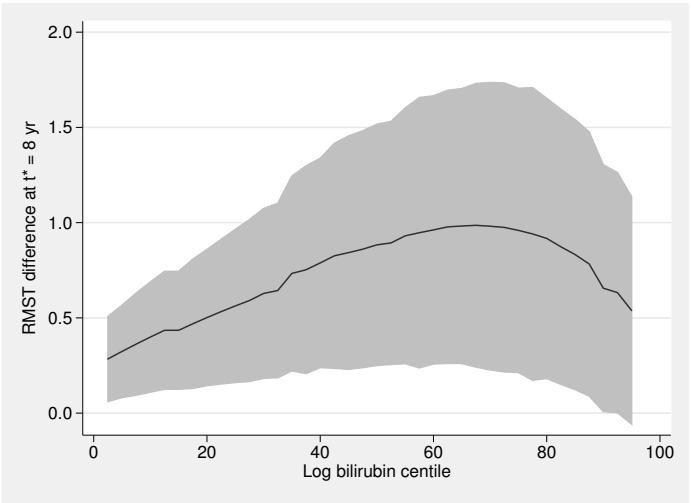

Figure 2. RMST difference at $t^* = 8$ years, with pointwise 95% CI, according to centiles of the distribution of log bilirubin. Estimates are from the $(3, 1)$ d.f. model, each conditional on one value of log bilirubin, and are adjusted for age and albumin.

Patients with larger bilirubin values (who have worse survival prospects) gain larger survival increments from azothiaprine. However, the treatment effect peaks at about the 70th centile and declines for larger values. This is due to the "floor effect", as illustrated in figure 3.

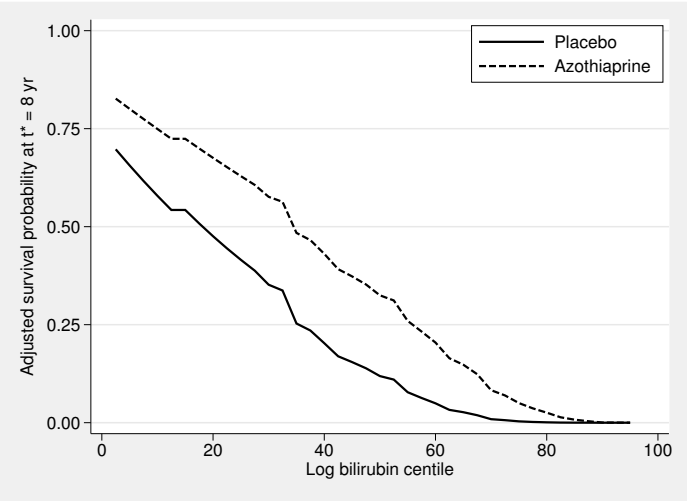

Figure 3. Predicted survival probability at $t^* = 8$ years in each treatment group, according to centiles of the distribution of log bilirubin. Estimates are from the $(3, 1)$ d.f. model and are adjusted for age and albumin.

The survival difference at 8 years narrows as log bilirubin increases above about the 50th centile. It follows that the values of the integrated survival curves (that is, the RMST) also get closer together, accounting for the smaller treatment effect. For patients with bilirubin above the 70th centile, the prognosis is so poor that by 8 years, there is little survival time left for azothiaprine to "buy". We would expect a less extreme result at earlier times, when there is more survival time to "buy", and this is precisely what occurs (data not shown). At $t^* = 4$ years, for example, the treatment-effect curve peaks at the 90th centile of log bilirubin.

This example illustrates some of the complexities that an appraisal of a treatment effect fuller than one HR can provide. Appropriate handling of the time domain in such trials is crucial for obtaining an adequate understanding of the quantitative effect of treatment on survival and its relation to prognosis. Conditional estimation may play an important role here.

4.4 Plotting survival curves

To calculate the RMST using the methods described above, one must estimate the unadjusted or directly adjusted survival curves. The curves are then integrated on $[0, t^*]$ to give RMST values. The survival curves are themselves informative.

Figure 4 compares the Kaplan–Meier estimates with unadjusted estimates from the $(3, 1)$ d.f. FPM, which includes a term for the nonproportional hazard of the treatment effect.

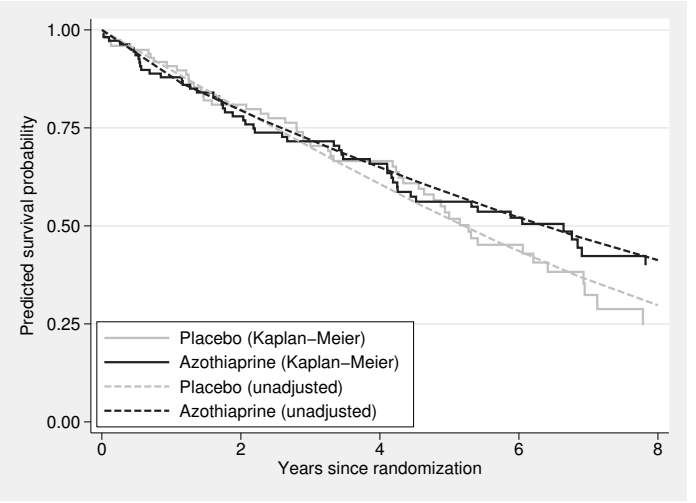

Figure 4. Survival curves estimated by the Kaplan–Meier method (solid lines) and by a $(3, 1)$ d.f. FPM (broken lines), not adjusted for prognostic factors

The shape of the "observed" survival curves is well approximated by the FPM curves, which cross about two years after randomization. After that, survival on azothiaprine appears slightly better than placebo.

In figure 5, a different picture is obtained with adjustment.

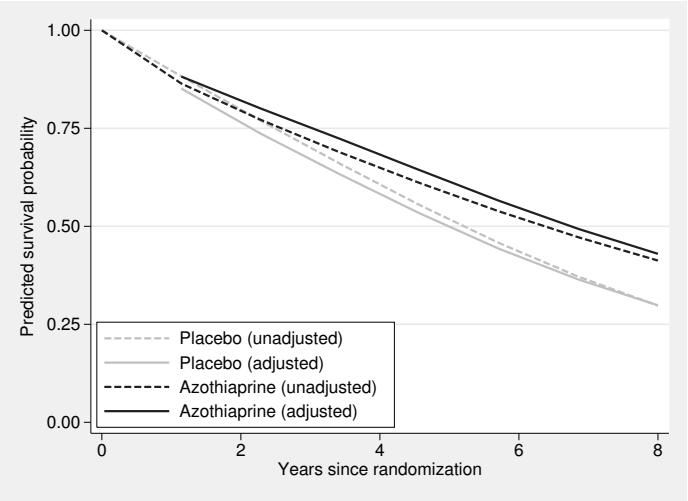

Figure 5. Survival curves estimated by $(3, 1)$ d.f. FPMs: solid lines, adjusted for prognostic factors; broken lines, unadjusted for prognostic factors

The adjusted survival curve for azothiaprine dominates that for placebo for all $t > 0$. Because we have not altered the parts of the model representing the treatment effect, the change is due only to adjustment. The separation between the unadjusted curves "catches up" with the adjusted curves by about $t = 8$ years.

4.5 Sensitivity of the RMST estimates to model specification

Because the estimates of the RMST are model dependent, one may wish to investigate how stable the estimates are with respect to possible variations in the FPM. Here we perturb the d.f. for the baseline spline (d.f.$_b$) and for the treatment-time interaction spline (d.f.$_t$). Recall that the prespecified model has (d.f.$_b$, d.f.$_t$) $= (3, 1)$. We also assess model fit in terms of the Akaike information criterion (AIC) and the Bayesian information criterion statistics for the fitted FPMs.

Figure 6 shows $\Delta(8)$ plotted against d.f.$_t$, the time-dependent d.f.

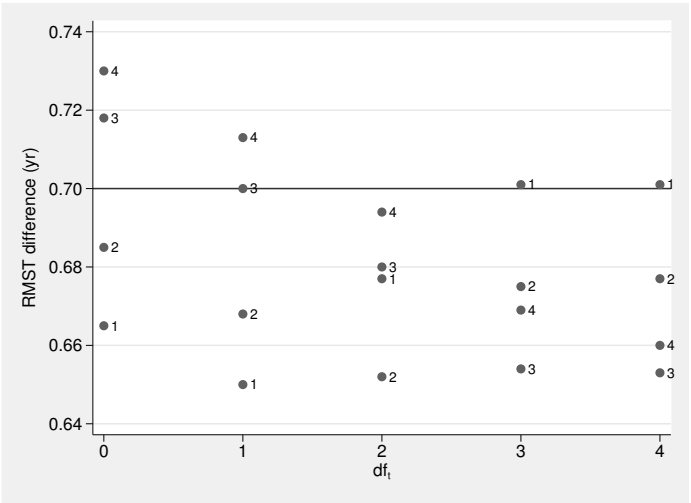

Figure 6. RMST difference at $t^* = 8$ years estimated by various more or less complex FPMs, (d.f.$_b$, d.f.$_t$) d.f., with differing values of d.f.$_b$ and d.f.$_t$. All models are adjusted for three prognostic factors. Labels affixed to points denote values of d.f.$_b$. The x axis shows d.f.$_t$. The horizontal line shows the default estimate. The four models with d.f.$_t = 0$ have PH of the treatment effect.

The labels affixed to the points denote d.f.$_b$, the baseline spline d.f. There is some variation in the estimate of $\Delta(8)$, depending on the precise model that is used. The default model (d.f.$_b = 3$, d.f.$_t = 1$) gives an estimate that is somewhat higher than most of the other models. However, the CI for each estimate is rather wide—$(0.17, 1.22)$ for the $(3, 1)$ d.f. model. A strategy to incorporate additional, between-model variation in the estimate of $\Delta(8)$ and its CI would be to use model uncertainty techniques (for example, see Buckland, Burnham, and Augustin [1997]). If we apply Buckland's method with AIC weights, we obtain $\Delta(8) = 0.695$ (SE 0.273) years, nearly identical to the default estimate of $\Delta(8) = 0.700$ (SE 0.268). This sensitivity analysis does not raise concerns about use of the default $(3, 1)$ d.f. FPM here.

According to minimum AIC, the best-fitting model is the PH model (d.f.$_b = 3$, d.f.$_t = 0$), for which AIC = 399.52. The default model is the second best (AIC = 400.56).

We conclude that allowing for model uncertainty hardly changes the estimated treatment effect or its uncertainty, which gives some confidence in the default values.

5 The strmst command

5.1 Syntax

The syntax of `strmst` is as follows:

```
strmst trt_varlist [if] [in], tstar(#) [adjust(adj_varlist) at(atlist)
    bootopts(options) bootstrap(#) df(#) dftvc(#) nint(#) nt(#)
    saving(filename[, replace]) scale(scalename) showsettings
    surv(surv [t]) ]
```

Important note: Before `strmst` can be used, `stpm2` must first be installed. Version 1.5.4 or later of `stpm2` is required.

The options for `strmst` are described below.

5.2 Description

`strmst` computes the RMST for each member of *trt_varlist*, which is a set of one or more binary dummy variables indicating treatment groups in a randomized controlled trial with a time-to-event outcome variable. The RMST is estimated at a prespecified time horizon, t^*, within a suitable FPM, which includes at least the dummy variables in *trt_varlist*.

By default, the FPM that is fit includes time-dependent treatment effects. Thus proportionality of the treatment effects on the scale of the model is not assumed. In particular, the model can accommodate nonproportional hazards of the treatment effects. Proportionality for the treatment variables can be imposed by specifying the `dftvc(0)` option.

`strmst` also calculates the difference in the RMST between each treatment group (as coded in *trt_varlist*) and the lowest level (all dummy variables in *trt_varlist* set to 0), which represents the control or reference group of the trial. The difference is an estimate of the treatment effect for each group compared with the control.

The analyst can avoid adjustment by omitting the `adjust()` option or adjust for covariates in *adj_varlist* in the FPM. `strmst` calculates and presents SEs for RMST values and their differences by using the delta method and, optionally, bootstrap resampling.

If `adjust()` is supplied, the RMST estimates are adjusted for covariates in *adj_varlist*. By default, proportionality on the scale defined by `scale()` is assumed to be PH. The FPM used to estimate survival curves in the various treatment groups includes treatment dummies from *trt_varlist* and the members of *adj_varlist*. The method known as direct adjustment is used; see section 9.3 of Royston and Lambert (2011) for an explanation of the method, which, essentially, involves marginalizing survival curves over the distribution of observed covariates.

Note that factor variables are not allowed in either *trt_varlist* or *adj_varlist*. However, the `xi:` command may be used in conjunction with the usual prefix `i.` to convert categorical variables to corresponding dummy variables on the fly.

5.3 Options

`tstar(#)` specifies the time horizon at which the RMST is to be calculated. This time is typically near the end of the follow-up period of the trial. `tstar()` is required.

`adjust(`*adj_varlist*`)` specifies adjustment of estimates for variables in *adj_varlist*. These may be a mixture of binary and continuous variables.

`at(`*atlist*`)` computes the RMST fixing values of covariates in *adj_varlist* according to *atlist*, which has syntax *varname* # [*varname* # ...]. For example, `at(x1 1 x3 50)` would evaluate the RMST at `x1` = 1 and `x3` = 50. Note that `x1` and `x3` must appear in *adj_varlist*. Note also that the use of `at()` does not alter the model being fit. The model fit is not restricted to the subset implied by the `at()` condition. Default behavior in the absence of `at()` is to marginalize survival curves and hence the RMST values over the observed distribution of covariates in *adj_varlist*.

`bootopts(`*options*`)` specifies options of the `bootstrap` command. It applies only if the `bootstrap()` option is used.

`bootstrap(#)` provides bootstrap estimates of SEs and confidence limits in addition to those from the default delta method. # specifies the number of bootstrap replications to be created. A reasonable suggestion for # is 200, but more than 200 can be used for more precise reproducibility of the estimated SE. The default is `bootstrap(0)`, meaning no bootstrap calculations are done.

`df(#)` specifies the degrees of freedom for the baseline spline function in the flexible parametric survival model to be used to estimate survival functions and hence the RMST. The default is `df(3)`.

`dftvc(#)` specifies the degrees of freedom for spline functions for time-dependent treatment effects in the FPM. If # is set to 1 or more, a time-dependent treatment effect is included that is increasingly complex as # increases. If # is set to 0 or less, no time dependency of the treatment effect is included—that is, proportionality of the treatment effects on the chosen scale is imposed. For `scale(hazard)` models (the default), the result with # = 0 is a PH model. The default is `dftvc(1)`, meaning a simple type of nonproportional hazards model is fit.

`nint(#)` specifies the number of integration time points for calculating the RMST from estimated survival curves. This option is rarely needed. The minimum # is 10, and the default is `nint(1001)`.

`nt(#)` specifies the number of time points for calculating the estimated survival curves. The corresponding times are created on an equally spaced grid of # values from 0 to `tstar()` inclusive. The minimum # is 2. There is no default. `nt()` requires `saving()` also to be specified.

`saving(`*filename*[`, replace`]`)` saves the predicted (adjusted) survival curves and corresponding times on the interval $[0, t^*]$ to a file called *filename*`.dta`. The survival curves are stored under default variable names `surv1`, `surv2`, etc., corresponding to the ordering of the treatment variable. When *filename*`.dta` is merged with the original data, the combined contents can be useful for plotting and comparing with unadjusted predicted or Kaplan–Meier curves. `saving()` requires `nt()` also to be specified. See also the `nt()` option.

`scale(`*scalename*`)` specifies the scale on which the FPM is to be fit. The default is `scale(hazard)`.

`scale(hazard)` fits a model on the log cumulative-hazard scale, that is, the scale of $\ln\{-\ln S(t)\}$. If no time-dependent effects are specified, the resulting model has PHs.

`scale(odds)` fits a model on the log cumulative-odds scale, that is, $\ln[\{1 - S(t)\}/S(t)]$. If no time-dependent effects are specified, the resulting model has proportional odds.

`scale(normal)` fits a model on the normal equivalent deviate scale, that is, a probit link for the survival function, `invnormal(`$1 - S(t)$`)`. If no time-dependent effects are specified, the result is a type of probit model.

`scale(theta)` fits a model on a scale defined by the value of theta for the Aranda-Ordaz family of link functions, that is, $\ln[S(t)^{-\texttt{theta}} - 1/\texttt{theta}]$. Note that `theta` $= 1$ corresponds to a proportional odds model and `theta` $= 0$ to a proportional cumulative hazards model.

`showsettings` displays the type of model being fit and whether or not proportionality of the treatment effects on the given scale is assumed.

`surv(`*surv* [*t*]`)` defines names for variables containing predicted survival probabilities and times stored with the `saving()` option. For example, if there were three treatment groups, there would correspondingly be three survival curves, and these would be stored in file *filename*`.dta` with the variable names *surv*`1`, *surv*`2`, and *surv*`3`. The time variable would be stored in the same file with name *t*. The defaults are `surv` for *surv* and `t` for *t*. `surv()` applies only when `saving()` has been specified. Otherwise, it is ignored.

5.4 Examples

```
. sysuse cancer
. tabulate drug, generate(d)
. strmst d2 d3, tstar(20)
. xi: strmst i.drug, tstar(20)
. xi: strmst i.drug, adjust(age) tstar(20) at(age 49)
. xi: strmst i.drug, adjust(age) tstar(20) dftvc(0) showsettings
. recode age 47/54=1  55/59=2 60/67=3, gen(ageg)
. xi: strmst i.drug, adjust(i.ageg) tstar(20)
. xi: strmst i.drug, adjust(age) tstar(20) dftvc(0) bootstrap(500)
```

```
. xi: strmst i.drug, adjust(age) tstar(30) df(2) saving(surv1) surv(s_adj t_adj)
. xi: strmst i.drug, tstar(30) df(2) saving(surv2) surv(s_unadj t_unadj)
```

6 Comments

I have described the definition, derivation, estimation, and use of RMST, principally in the context of randomized controlled trials (for example, in cancer). However, RMST is a potentially useful summary statistic in any study with a time-to-event outcome. Such an application is prognostic modeling, in which the aim is to predict the clinical course of a disease or other condition from the "prognostic factors" measured at baseline ($t = 0$). Chapter 6 of Royston and Lambert (2011) gives some examples. RMST can be calculated for any type of prediction within this context by specifying the `rmst` option of `predict` following model estimation with `stpm2`.

A rich set of standard parametric survival models has been implemented through the official Stata command `streg`. Implementation of RMST for such models has not been done, nor has it been included in `strmst`. It would require out-of-sample prediction of suitable survival curves, as needed for numerical integration of the survival probabilities over a fine, equally spaced grid of times covering $(0, t^*)$. To provide direct adjustment for covariates, one would need to average predicted survival curves over covariate patterns. Because it is not a standard feature of `predict` following `streg`, averaging would need to be coded specifically. As already noted, however, three standard parametric models (Weibull, loglogistic, and lognormal) are already available with `stmrst` or `stpm2` using the `df(1)` option. More "exotic" variants, such as frailty models, have not (yet) been implemented in `stpm2`.

One point about the role of t^* may need clarification. One might think that by selecting a time horizon t^*, we effectively discard or right-censor all events occurring after t^*, which therefore cannot impact the RMST. This is a misunderstanding; such a procedure would indeed waste information. We fit the FPM using all the data and then compute the RMST for a selected t^* based on the parameter estimates. In this way, all observations contribute to the estimates, but the FPM need be fit only once. The resulting parameter estimates can be used for any desired t^* within the range of the observed times.

I believe that the `strmst` program will help statisticians working in clinical trials to realize that RMST is a viable and useful measure of the treatment effect. RMST is also available for more general settings with the `predict, rmst` command following model estimation with `stpm2`. A detailed account of the theory and applications of FPMs is given by Royston and Lambert (2011).

7 Acknowledgments

I thank Dr. Tim Morris, Dr. Gordana Jovič, and an anonymous reviewer for insightful comments that helped to improve the software and the article. The derivation of RMST

from first principles as an integrated survival function was kindly provided by Dr. Erik Parner.

8 References

Andersen, P. K., M. G. Hansen, and J. P. Klein. 2004. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* 10: 335–350.

Andersen, P. K., and M. Pohar Perme. 2010. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 19: 71–99.

Andersson, T. M.-L., and P. C. Lambert. 2012. Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models. *Stata Journal* 12: 623–638.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: An integral part of inference. *Biometrics* 53: 603–618.

Christensen, E., J. Neuberger, J. Crowe, D. G. Altman, H. Popper, B. Portmann, D. Doniach, L. Ranek, N. Tygstrup, and R. Williams. 1985. Beneficial effect of azathioprine and prediction of prognosis in primary biliary cirrhosis: Final results of an international trial. *Gastroenterology* 89: 1084–1091.

Irwin, J. O. 1949. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene* 47: 188–189.

Lambert, P. C., and P. Royston. 2009. Further development of flexible parametric models for survival analysis. *Stata Journal* 9: 265–290.

Overgaard, M., P. K. Andersen, and E. T. Parner. 2015. Regression analysis of censored data using pseudo-observations: An update. *Stata Journal* 15: 809–821.

Parner, E. T., and P. K. Andersen. 2010. Regression analysis of censored data using pseudo-observations. *Stata Journal* 10: 408–422.

Royston, P., D. G. Altman, and W. Sauerbrei. 2006. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine* 25: 127–141.

Royston, P., and P. C. Lambert. 2011. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. College Station, TX: Stata Press.

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

———. 2011. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* 30: 2409–2421.

———. 2013. 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 Medical Research Methodology* 13: 152.

Yusuf, S., D. Zucker, E. Passamani, P. Peduzzi, T. Takaro, L. D. Fisher, J. W. Kennedy, K. Davis, T. Killip, R. Norris, C. Morris, V. Mathur, E. Varnauskas, and T. C. Chalmers. 1994. Effect of coronary artery bypass graft surgery on survival: Overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet* 344: 563–570.

About the author

Patrick Royston is a medical statistician with more than 30 years of experience, with a strong interest in biostatistical methods and in statistical computing and algorithms. He works largely in methodological issues in the design and analysis of clinical trials and observational studies. He is currently focusing on alternative outcome measures in trials with a time-to-event outcome; on problems of model building and validation with survival data, including prognostic factor studies and treatment-covariate interactions; on parametric modeling of survival data; and on novel clinical trial designs.