The Stata Journal (2017)
17, Number 2, pp. 405–421

A combined test for a generalized treatment effect in clinical trials with a time-to-event outcome

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

Abstract. Most randomized controlled trials with a time-to-event outcome are designed and analyzed assuming proportional hazards of the treatment effect. The sample-size calculation is based on a log-rank test or the equivalent Cox test. Nonproportional hazards are seen increasingly in trials and are recognized as a potential threat to the power of the log-rank test. To address the issue, Royston and Parmar (2016, *BMC Medical Research Methodology* 16: 16) devised a new "combined test" of the global null hypothesis of identical survival curves in each trial arm. The test, which combines the conventional Cox test with a new formulation, is based on the maximal standardized difference in restricted mean survival time (RMST) between the arms. The test statistic is based on evaluations of RMST over several preselected time points. The combined test involves the minimum p-value across the Cox and RMST-based tests, appropriately standardized to have the correct null distribution. In this article, I outline the combined test and introduce a command, `stctest`, that implements the combined test. I point the way to additional tools currently under development for power and sample-size calculation for the combined test.

Keywords: st0479, stctest, randomized controlled trial, time-to-event outcome, restricted mean survival time, treatment effect, hypothesis testing, flexible parametric model, jackknife

1 Introduction

Most randomized controlled trials with a time-to-event outcome are designed and analyzed assuming proportional hazards (PH) of the treatment effect. The sample-size calculation is based on a log-rank test or the equivalent Cox test. However, nonproportional hazards (non-PH) are increasingly recognized as an issue (for example, Trinquart et al. [2016]). Significant non-PH may be present in about a quarter of cancer trials (Trinquart et al. 2016). Nonstatistically significant, but still practically important non-PH are likely to be present in a much larger proportion of trials, particularly as trial sample size and follow-up time tend to increase, conferring higher power to detect non-PH.

Possible reasons for non-PH include treatments that really do have time-dependent effects. For example, a research treatment given only over a limited period may be effective early but wear off later. Alternatively, a treatment may have no effect for a relatively long period after randomization but "kick in" further on, which is a "late effect" of a type sometimes seen in prevention trials and screening trials. Treatments with different modes of action, such as surgery, drug treatment, and watchful waiting, may induce non-PH because of dissimilarity between the shapes of the hazard functions in the control and research arms. Additionally, the presence of a subpopulation with a differential response to the research treatment may distort the survival curve.

Concerns about use of the hazard ratio (HR) as a summary measure and as the basis of a test of the treatment effect in such trials include poor interpretability and potential loss of power of the associated Cox or log-rank test. Difference (or ratio) in restricted mean survival time (RMST) between treatment groups is gaining popularity as a summary measure and as the basis of a possible test of a treatment effect. RMST at some time point ($t^* > 0$) is the integral of the survival function at t^*, that is, the "area under the survival curve" from 0 to t^*. It is interpreted as the mean of the survival-time distribution truncated at t^*. The difference, ΔRMST, defined as RMST in a research arm minus RMST in the control arm, is the integrated difference between the survival functions—in other words, the (signed) area between the survival curves up to t^*. Conventionally, a "large" positive value of ΔRMST is regarded as a "good" trial outcome because it represents an extension of survival time because of the research regimen, at least up to t^*. Further details and an implementation of RMST and ΔRMST in the user-written `strmst` command may be found in Royston (2015); also see the `strmst2` command (Cronin, Tian, and Uno 2016).

One might surmise that, with a suitable choice of t^*, ΔRMST divided by its standard error (SE) might provide a useful test statistic for the "global" null hypothesis $H_0 : S_0(t) = S_1(t)$ for any $t > 0$, where $S_j(t)$ is the survival function in the jth group ($j = 0, 1$) with $j = 0$ denoting the control group. The problem is the choice of t^*. A single value is fragile regarding power. To protect power, one would prefer to test over a range of t^* values. Recognizing such a requirement, Royston and Parmar (2016) proposed a test of H_0 based on evaluating the maximal chi-squared statistic $C_{\max} = \max\left(Z^2\right)$ over several time points, where $Z = \Delta\text{RMST}/\text{SE}(\Delta\text{RMST})$. Arguing pragmatically, Royston and Parmar (2016) determined $C_{\max}$ over 10 equally spaced values of t^* between the 30th and 100th centiles of the failure times in the dataset. Starting with $C_{\max}$, they developed an approach to testing H_0 that they called the "combined test".

My principal aim here is to present a new command, `stctest`, that implements the combined test. In section 2, I outline the methodological steps leading to the combined test and describe different "flavors" of the test. In section 3, I present the `stctest` command. In section 4, I apply the methodology to an example trial dataset. Section 5 is a discussion.

2 Methods

2.1 Estimation of RMST

Estimation of RMST at some time t^* requires determining the area under the survival curve from 0 to t^*. I consider two methods: i) `ps`, using jackknife estimation of pseudovalues (Andersen, Hansen, and Klein 2004), equivalent to integrating the Kaplan–Meier curve; and ii) `rp`, integration of smooth survival curves predicted from flexible parametric survival models (Royston and Parmar 2002; Lambert and Royston 2009; Royston and Lambert 2011), also known as Royston–Parmar (RP) models. Next, I briefly describe these methods.

Jackknife estimation from pseudovalues

Andersen, Hansen, and Klein (2004) described the use of "pseudo-observations" (I call them pseudovalues) to provide nonparametric estimates of RMST at the individual participant level. Pseudovalues are leave-one-out (jackknife) estimates of a parameter of interest, here the RMST, constructed in such a way that their sample mean estimates the RMST. They are computed from the Kaplan–Meier estimate of the survival curve for the sample. The effects of covariates on the RMST may be modeled with the pseudovalues as the response variable in generalized linear models with a suitable link function. Standard errors of parameter estimates use the robust "sandwich" estimator in Stata through the `robust` estimation option. Because pseudovalues are based on Kaplan–Meier estimates, they are distribution free.

In Stata, pseudovalues for RMST are available through the user-written `stpmean` command (Parner and Andersen 2010; Overgaard, Andersen, and Parner 2015). A treatment effect can be estimated by a command of the form `regress` *psvar trtvar*`, robust`. The response variable, *psvar*, contains the pseudovalues for some t^*, as estimated by `stpmean`. The regression coefficient for *trtvar* estimates the arithmetic difference in RMST between the treatment groups.

RP models

Conceptually, RP models fit the baseline distribution function explicitly using a suitable smoother; Royston and Parmar (2002) chose restricted cubic spline functions. Effects of covariates $\mathbf{x}$ are accommodated in generalized linear models of the form

$$g_\theta \{S(t;\mathbf{x})\} = g_\theta \{S_0(t)\} + \mathbf{x}\boldsymbol{\beta}$$

where $S(t;\mathbf{x})$ and $S_0(t)$ are the survival and baseline survival functions, respectively, and $g_\theta(\cdot)$ is a monotonic link function. See Royston and Lambert (2011, 118–119) for further details of this class of models.

Here I use the subclass with a complementary log-log link function, defined by $g_\theta \{S(t;\mathbf{x})\} = \ln\{-\ln S(t;\mathbf{x})\} = \ln\{H(t;\mathbf{x})\}$, the log cumulative-hazard function:

$$\ln H(t;\mathbf{x}) = \ln H_0(t) + \mathbf{x}\boldsymbol{\beta}$$

The function $\ln H_0(t)$ is modeled using restricted cubic splines in $\ln t$, the complexity of which is determined by the number and position of user-selected interior knots (polynomial join points). If the covariate effects $\boldsymbol{\beta}$ are independent of time, as in the above expression, the formulation gives a parametric PH model.

RP models are easily extended to include non-PH covariate effects; see Royston and Lambert (2011, sect. 7.6). In a randomized controlled trials context, time-dependent effects are achieved by fitting different spline functions in each treatment group. Depending on the degree of freedom (d.f.) chosen for the spline functions, the approach potentially provides sufficient flexibility to represent many varieties of non-PH patterns. Here I suggest using a relatively complex spline model with five d.f. (equivalent to four internal knots) in each treatment group, providing estimates of the treatment effect that are comparably flexible with those from the method based on pseudovalues.

In general, the tool recommended for fitting RP models in Stata is `stpm2` (Lambert and Royston 2009). Estimation of RMST after fitting an RP model with `stpm2` is straightforward. One uses the `rmst` option of `predict` together with a second option, `tmax(#)`, to define t^*. Standard errors and confidence intervals (CIs) are supported through the `stdp` and `ci` options. Further useful options are `at()` to predict RMST at specific values of covariates and `zeros` to predict at baseline (all covariates set to zero). For applications to trials, please see the user-written `strmst` command (Royston 2015), which conveniently packages RMST calculations from RP models.

2.2 Approximate combined test

The motivation for $C_{\max}$ (defined in the *Introduction*) as the basis of a test of the treatment effect is to try to identify the largest standardized treatment effect over a relevant time interval. Because of multiple testing at 10 time points, the null distribution of $C_{\max}$ is not central chi-squared on 1 d.f. To correct for multiplicity and arrive at a usable test statistic, Royston and Parmar (2016) took the following steps:

1. To estimate a p-value associated with $C_{\max}$, they first create M values of $C_{\max}$ in the null case. This is done by randomly permuting the treatment label in the given dataset, thereby "scrambling" any treatment-outcome association. $C_{\max}$ is calculated in each permuted sample; call the resulting values $C_1, \ldots, C_M$. Suppose that $r \geq 0$ of the C_i exceeds $C_{\max}$. The larger r is, the weaker the evidence that $C_{\max}$ is "extreme" and the larger the corresponding p-value. Their continuity-corrected estimate of the p-value is $p_{\text{perm}} = \{r + (1/2)\}/(M+1)$. (Note that r can take any of the $M+1$ values $0, 1, \ldots, M$.) The smallest p_{perm} that can result with a given M is $1/(2M+2)$. A binomial-based exact CI for r/M may be used to calculate a CI for p_{perm}.

2. Such a permutation test has a stochastic element. Let $p_{\max}$ be the p-value corresponding to $C_{\max}$ according to a chi-squared distribution on one d.f. In Stata terms, $p_{\max}$ = `chi2tail(1,`$C_{\max}$`)`. To stabilize the test, they used simulations based on several real datasets to derive $\widetilde{p}_{\text{perm}}$, an empirical approximation to p_{perm} as a function of $p_{\max}$,

$$\widetilde{p}_{\text{perm}} = 1.762\,(p_{\max})^{0.885} - 0.802\,(p_{\max})^{2.547}$$

3. Next, they developed a new test combining $\widetilde{p}_{\text{perm}}$ with the Cox test p-value, p_{Cox}. The aim was to capitalize on the strengths of each test across a range of patterns of survival curves, including PH and non-PH examples. They defined

$$p_{\min} = \min\left(p_{\text{Cox}}, \widetilde{p}_{\text{perm}}\right)$$

Although the individual null distributions of p_{Cox} and $\widetilde{p}_{\text{perm}}$ are approximately uniform on $(0, 1)$, p_{Cox} and $\widetilde{p}_{\text{perm}}$ are correlated, and the null distribution of $p_{\min}$ is not expected to be uniform. They approximated the null distribution of $p_{\min}$ empirically using a two-parameter beta distribution. To calculate an approximate p-value, $\widetilde{p}_{CT}$, from a given $p_{\min}$, they applied the formula (in Stata terms) $\widetilde{p}_{CT}$ = `ibeta(`$a, b, p_{\min}$`)`, where `ibeta(`a, b, x`)` is the incomplete beta function with parameters a, b and argument x $(0 < x < 1)$. For the two-sided test, they estimated $a = 1$, $b = 1.5$.

Royston and Parmar (2016) did not provide an expression for $\widetilde{p}_{CT}$ for use in one-sided tests. However, in subsequent work using similarly constructed simulations, they obtained the following two-parameter beta approximation to the null distribution of $p_{\min}$ for the one-sided test: $a = 0.9642$, $b = 1.2581$.

2.3 Permutation combined test

Description

In an analysis of simulations based on data from 20 selected randomized trials, Royston and Parmar (2016) showed that $\widetilde{p}_{CT}$ maintained approximately the correct significance level in the null case of no treatment effect. However, the possibility of heterogeneity remained in other (unconsidered) trials, meaning that $\widetilde{p}_{CT}$ might be (slightly) too large or (slightly) too small in some trials. In critical cases, this might matter. Ensuring the integrity of a p-value for a treatment effect in a randomized trial is important.

With such a motivation, I extend the permutation approach used with $C_{\max}$ to create a permutation-based combined test, as follows:

1. Determine $C_{\max}$, $p_{\max}$, $\widetilde{p}_{\text{perm}}$ (but not p_{perm}), and p_{Cox} on the original dataset, as described above. Note that none of these quantities is stochastic. Let $p_{\min}^{\text{orig}} = \min\left(p_{\text{Cox}}, \widetilde{p}_{\text{perm}}\right)$.
2. Determining a permutation p-value, p_{CT}, for the combined test rests on assessing the relative position of $p_{\min}^{\text{orig}}$ in the null (permutation) distribution of $p_{\min}$. The method is similar to the determination of p_{perm} given above.
3. In each of M samples with a random permutation of the treatment label, determine $p_{\max}$, $\widetilde{p}_{\text{perm}}$, p_{Cox}, and hence $p_{\min}$, thus establishing a sample of size M from the permutation distribution of $p_{\min}$.

4. Calculate the permutation combined test as $p_{CT} = \{r + (1/2)\}/(M+1)$, where r is the number of samples in which $p_{\min}$ is smaller (that is, "more significant") than $p_{\min}^{\text{orig}}$. A CI for p_{CT} may be found via a binomial based CI for r/M and some simple algebra.

Validation of type 1 errors

If the nonstochastic combined test $\widetilde{p}_{CT}$ has the correct type 1 error probability, the expected proportion of M samples with a random permutation of the treatment label in which $\widetilde{p}_{CT} < \alpha$ should be α for any trial and choice of α. Here α is interpreted as the nominal significance level and is the appropriate critical value for the test. I tested this important characteristic through a heterogeneity chi-squared statistic, defined for a given α by $C_{H;\alpha} = \sum_{j=1}^{20} (p_{j;\alpha} - \alpha)^2 / \text{var}(p_{j;\alpha})$. For trial j, $p_{j;\alpha}$ denotes the proportion of M samples in which $\widetilde{p}_{CT} < \alpha$; and $\text{var}(p_{j;\alpha}) = p_{j;\alpha}(1 - p_{j;\alpha})/M$. If $E(p_{j;\alpha}) = \alpha$ for each j, then $C_{H;\alpha}$ is distributed approximately as central chi-squared on 20 d.f.

For each of 20 trials, I created $M = 5000$ permutation samples. I estimated $p_{\min}$ and hence $\widetilde{p}_{CT} = \texttt{ibeta}(a, b, p_{\min})$ in each sample, thus generating 100,000 observations of $\widetilde{p}_{CT}$ for the two-sided combined tests.

For conventional values $\alpha \in \{0.01, 0.025, 0.05, 0.1\}$, I estimated $p_{j;\alpha}$ as the proportion of $M = 5000$ values so that $\widetilde{p}_{CT} < \alpha$ in trial j $(j = 1, \ldots, 20)$. Aside from chance variation, the $p_{j;\alpha}$ should be consistent with α. Figure 1 shows the $p_{j;\alpha}$ with 95% CIs for the two-sided combined test.

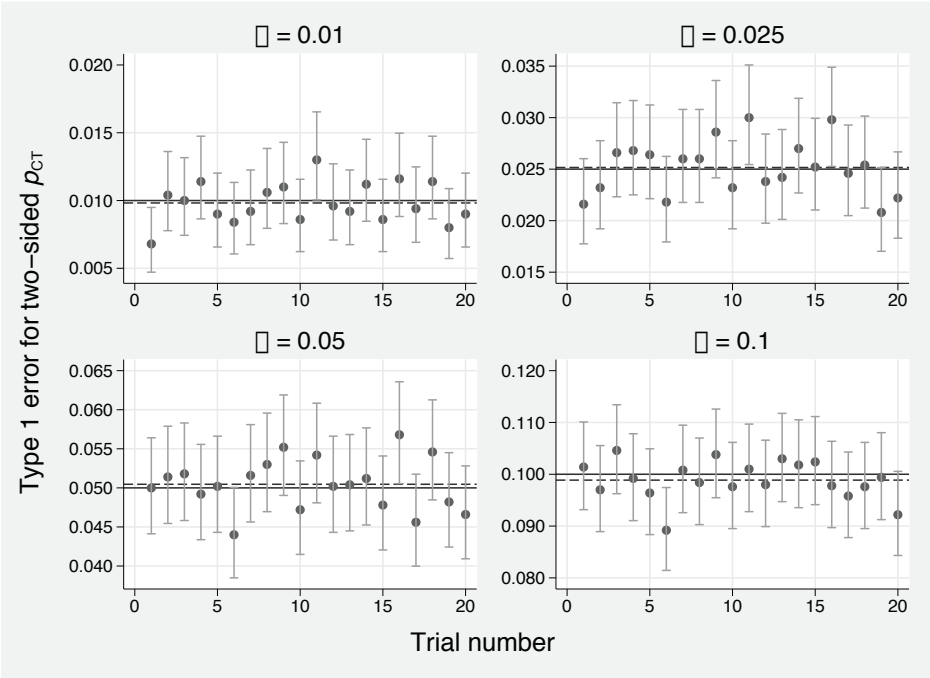

Figure 1. Empirical type 1 error probabilities $p_{1;\alpha}, \ldots, p_{20;\alpha}$ with 95% CIs for the two-sided combined test across 20 trials. Solid horizontal lines show critical values (α), and dashed lines show the corresponding p_α in the pooled dataset of 100,000 samples with a randomly permuted treatment label.

For each α, the $p_{j;\alpha}$ are scattered seemingly randomly around α and lie within about two SEs errors of α. For the pooled sample ($M = 100000$), p_α is close to α. The heterogeneity chi-squared ($C_{H;\alpha}$) is not significant at the 5% level for any of the four illustrated values of α.

Figure 2 shows an analogous plot for the one-sided combined test.

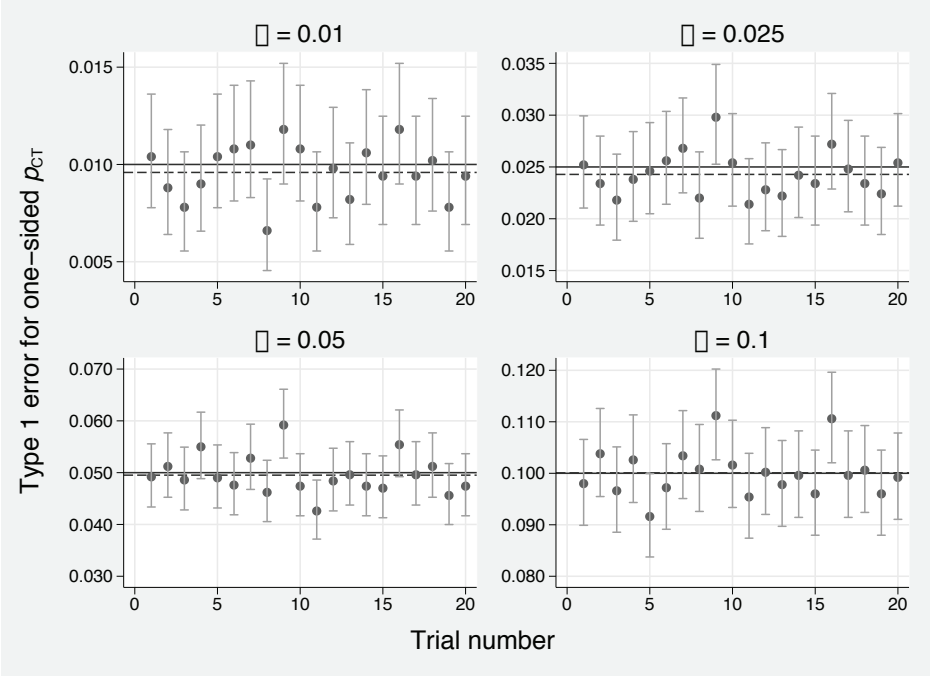

Figure 2. Empirical type 1 error probabilities $p_{1;\alpha}, \ldots, p_{20;\alpha}$ with 95% CIs for the one-sided combined test across 20 trials. Solid horizontal lines show critical values (α), and dashed lines show the corresponding p_α in the pooled dataset of 100,000 samples with a randomly permuted treatment label.

The values of $p_{j;\alpha}$ are again generally close to α. None of the four heterogeneity chi-squared is significant at the 5% level.

I conclude that the approximations that lead to $\widetilde{p}_{CT}$ for the two-sided and one-sided combined tests appear to work well. Nevertheless, the permutation combined test, p_{CT}, provides an important "safety net" for use in critical cases, for example, when $\widetilde{p}_{CT}$ is close to a critical value such as $\alpha = 0.05$.

3 The stctest command

The syntax of `stctest` is as follows:

`stctest` {`ps`|`rp`} *trt_varname* [*if*] [*in*] [`,` `adjust(`*adj_varlist*`)` `compare(`*#1 #2*`)` `detail` `df(`*#*`)` `dftvc(`*df_list*`)` `nperm(`*#*`)` `onesided(+|-)`]

Note that before all the features of `stctest` can be used, two programs must be installed: `stpmean` and `stpm2`. `stpmean` may be installed from the *Stata Journal* website using the commands

```
. net sj 15-3 st0202_1
. net install st0202_1
```

Also, `stpm2` may be installed or updated from the Statistical Software Components archive using the command

```
. ssc install stpm2, replace
```

Important: Please ensure you install (or update to) the most recent version of `stpm2`, as above.

3.1 Description

`stctest ps` and `stctest rp` carry out a combined significance test (Royston and Parmar 2016) of a generalized treatment effect comparing level *#1* of *trt_varname* with level *#2*. The data are assumed to arise from a randomized controlled trial with a time-to-event outcome and assumed to have been `stset`. Usually, *#1* denotes the control arm and *#2* a research arm. *trt_varname* may contain more than two levels (treatment arms), but `stctest` compares only selected pairs of levels as determined by the `compare(`*#1 #2*`)` option. Typically, a research treatment (a novel regimen under investigation) is compared with a control arm (standard of care or some other kind of reference therapy such as a placebo).

The combined test combines a standard log-rank or Cox test (implemented through `stcox` *trt_varname*) with a statistic derived from the maximal squared standardized between-arm difference in time-dependent RMST. Further details are given above and in the help file under *Remarks*.

`stctest ps` carries out the combined test "nonparametrically" using RMST "pseudovalues" calculated by the user-written `stpmean` command (Parner and Andersen 2010; Overgaard, Andersen, and Parner 2015). Pseudovalues are jackknife quantities derived from the Kaplan–Meier survival function and constructed so that their arithmetic mean estimates the RMST at a given time point, t^*.

`stctest rp` carries out the combined test "parametrically" using estimates of RMST derived from an RP model (Royston and Parmar 2002; Royston and Lambert 2011) fit by `stpm2` (Lambert and Royston 2009). Regression parameters are defined on the scale of the log cumulative-hazard function. To allow for the possibility of non-PH, the model includes a time-dependent treatment effect.

3.2 Options

I describe the more important options here. Lesser used options `df()` and `dftvc()` are described in the help file.

`adjust(`*adj_varlist*`)` adjusts RMST for variables in *adj_varlist*, allowed to be any mixture of binary, continuous, and factor variables. Note that `stctest ps` and `stctest rp` adjust differently for covariates. In both flavors, the Cox component of the combined test is a PH model that includes *trt_varname* and variables in *adj_varlist*. The RMST test component in `stctest ps` includes *trt_varname* and *adj_varlist* in multiple linear regression models for the RMST pseudovalues at the different time points. Thus `stctest ps` incorporates linear adjustment for covariates on the scale of RMST. In contrast, `stctest rp` includes *adj_varlist* in the hazards-scaled RP model that incorporates a time-dependent effect (non-PH) for treatment. Thus `stctest rp` adjusts linearly for covariates on the log cumulative-hazard scale, which, in the absence of time-dependent effects, is a PH model for these variables.

`compare(`*#1 #2*`)` selects the levels of the treatment variable to be compared. Usually *#1* denotes the control arm and *#2* a research arm. The default is `compare(0 1)`.

`detail` reports results of the component Cox and RMST tests in addition to p_{CT}, the p-value for the primary test (the combined test).

`nperm(#)` changes the mechanics of `stctest ps` and `stctest rp` so that the null distribution of the combined test statistic is derived directly from a permutation test procedure. Using the Stata `permute` command, the treatment covariate is randomly permuted # times, and the combined test is performed in each permuted dataset, providing multiple values of $\widetilde{p}_{\text{perm}}$, p_{Cox}, and $p_{\min}$. The ensemble constitutes a sample from the permutation distribution of $p_{\min}$. The relative position of $p_{\min}^{\text{orig}}$, the test statistic from the original data, in the permutation distribution of $p_{\min}$ estimates p_{CT} for the combined permutation test.

Using `nperm(#)` with $\# > 0$ allows one to estimate a p-value for the combined test that does not rely on empirical approximations. However, the variance of such a p-value may be large. If a p-value with a "narrow" CI is desired, a "large" value of # will be required, for example, 5,000 or more. Computation time increases linearly with #. Computation times with `stctest rp` will be particularly long, so the approach should be used only when absolutely needed.

The default is `nperm(0)`, meaning that the combined test p-value, p_{CT}, is obtained nonstochastically through a beta distribution approximation (see Royston and Parmar [2016]).

`onesided(+|-)` performs one-sided tests of the treatment effect. For the Cox test, one-sided p-values are reported. With `onesided(+)`, the direction of the test is that lower HRs have smaller p-values, because HR < 1 in most trials represents a "positive" test result. `onesided(-)` may be appropriate when the event of interest is a "good" outcome, for example, time to wound healing. With `onesided(+)`, the RMST test responds to RMST being higher in the research arm than the control arm, and vice versa for `onesided(-)`. In most trials, an increase in the mean time to event is a "good" outcome. The default is `onesided()`; that is, the option is unspecified, meaning that all tests are two sided.

4 Example

An interesting example is the PATCH1 trial of treatment for cellulitis of the leg, a common bacterial infection of the skin and underlying tissue (Thomas et al. 2013). In a prophylaxis phase, 274 patients were randomly assigned to placebo or treatment with penicillin. One of the main outcomes of interest was time to first disease recurrence during a no-intervention follow-up period. Only one event occurred after three years. For presentation purposes, follow-up time was truncated at three years. There were 128 first recurrences and 146 censored observations.

Figure 3 shows estimated "survival" curves by treatment group.

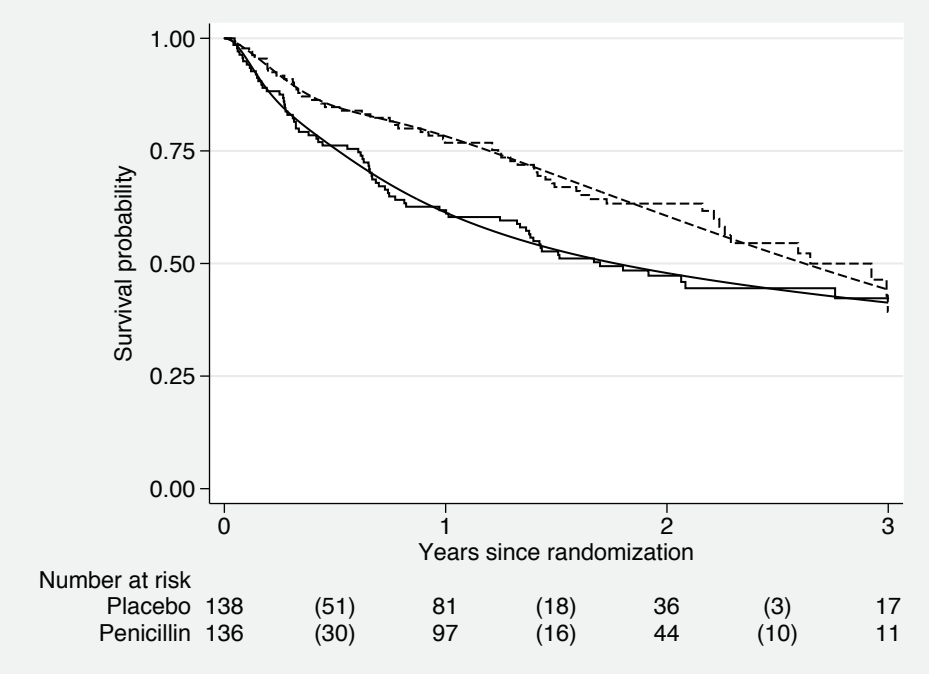

Figure 3. PATCH1 trial: Survival curves for time to first recurrence of cellulitis by treatment group. Unbroken lines, placebo group; dashed lines, penicillin group. Jagged lines, Kaplan–Meier curves; smooth curves, estimates from an RP model. Values in parentheses denote number of events in the corresponding time interval.

The Cox test of the treatment effect "just fails to achieve significance" at conventional levels, with $p = 0.052$. However, the Kaplan–Meier curves suggest a clear difference between treatments. The median time to recurrence of cellulitis increases by almost one year, from 1.70 years on placebo to 2.65 years on penicillin (difference = 0.95, SE = 0.41, $p = 0.021$). Applying the combined test with the `ps` (pseudovalues) method produces the following result. I include the `detail` option to see the component test results:

```
. use patch1
(PATCH1 trial (public release version), PR May 2016)
. stctest ps trt, detail
      Treatment |   Obs      p(CT)*
----------------+-----------------
       trt(0,1) |   274    0.023746
* Non-stochastic, from approximation to permutation test
p-values from tests underlying p(CT):
  p(Cox)    p(chi2)    p(perm)     p(min)
-------------------------------------------
0.051835   0.004893   0.015894   0.015894
```

The combined test is significant at the 0.05 level ($p = 0.02375$), similar to the test of medians. As discussed, the uncorrected p-value for the test of RMST differences is "too small" (0.00489). After correction, it is 0.01589, which is significant at the 0.05 level. `p(min)`, the smaller of the Cox and approximate permutation test p-values, is 0.01589. After adjustment for the null distribution of $p_{\min}$, the combined test p-value, `p(CT)`, is 0.02375.

Repeating the `stctest` command, but this time using the `nperm(5000)` option, gives the following output. To ensure reproducibility, I first set the random-number generator seed:

```
. set seed 123
. stctest ps trt, nperm(5000) detail
      Treatment |   Obs      p(CT)*  [95% Conf. Interval]
----------------+-----------------------------------------
       trt(0,1) |   274    0.025495   0.021314   0.030242
* Stochastic, from estimated permutation null distribution of p(min)
p-values from tests underlying p(CT):
  p(Cox)    p(chi2)    p(perm)     p(min)
-------------------------------------------
0.051835   0.004893   0.015894   0.015894
```

The p-value for the permutation version of the combined test is 0.02549, similar to the nonstochastic value of 0.02375. Displaying `return list` to see the stored quantities provides the following information:

```
. return list
scalars:
              r(nperm) =  5000
          r(delta_max) =  .2652538362205551
          r(tstar_max) =  2.103439807892
                r(pct) =  .025494901019796
             r(pct_ub) =  .030241944495036
             r(pct_lb) =  .021313527894152
               r(nsig) =  127
               r(pmin) =  .0158943253638914
              r(pperm) =  .0158943253638914
              r(pchi2) =  .0048928816887735
             r(pjoint) =  .021944650127736
                r(pgt) =  .0495280333667285
                r(plr) =  .0517262578749101
                 r(hr) =  .7080773912586498
               r(pcox) =  .0518346333672928
                 r(t2) =  2.997728890592487
                 r(t1) =  .3148614609571788
```

Of the $M =$ `r(nperm)` $= 5000$ permuted datasets, $p_{\min}$ is less than or equal to $p_{\min}^{\text{orig}} =$ `r(pmin)` $= 0.01589$ in $r =$ `r(nsig)` $= 127$ permuted datasets, giving $p_{CT} =$ `r(pct)` $= \{r + (1/2)\}/(M + 1) =$ (`r(nsig)` $+ 0.5$)/(`r(nperm)` $+ 1$) $= 128.5/5001 = 0.02549$.

Note that the 95% CI for `r(pct)` $= 0.02549$ is `r(pct_lb)` $= 0.02131$, `r(pct_ub)` $= 0.03024$. Although the CI is fairly wide, the upper bound is well below the reference level of 0.05, confirming that at conventional significance levels, there is a real effect of treatment.

A brief description of the remaining stored quantities is given in the help file. In particular, the Grambsch–Therneau test of the PH assumption, for which the p-value is returned in `r(pgt)` as 0.04953, is just significant at the 0.05 level. Non-PH may explain why the Cox test appears to have low power, despite the HR `r(hr)` $= 0.708$ being well below 1.0.

Rerunning the combined test using the `rp` method gives results that are similar but not identical to the `ps` method:

```
. stctest rp trt, detail
      Treatment |    Obs      p(CT)*
----------------+-----------------
       trt(0,1) |    274    0.017265
* Non-stochastic, from approximation to permutation test
p-values from tests underlying p(CT):
  p(Cox)    p(chi2)    p(perm)     p(min)
-------------------------------------------
0.051835   0.003409   0.011543   0.011543
```

Figure 4 illustrates time-dependent estimates of ΔRMST according to the `ps` and `rp` methods.

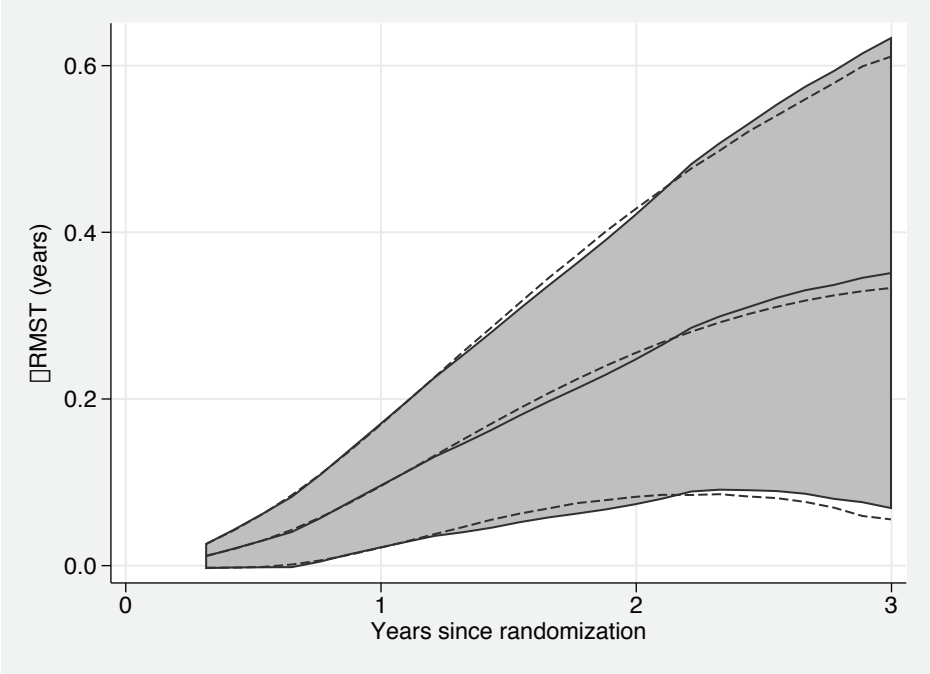

Figure 4. PATCH1 trial. Time-dependent estimates of ΔRMST, with pointwise 95% CIs, according to two methods. Solid lines and gray shaded area: `ps` method; dashed lines, `rp` method.

ΔRMST and pointwise 95% CIs were calculated at 25 equally spaced time points between the 30th and 100th centiles of the uncensored failure times (0.31 and 3.00 years, respectively). Note the considerable similarity of the two sets of estimates. At $t^* = 3.0$ years (for example), I find $\Delta\text{RMST} = 0.35\ (0.07, 0.63)$ years with `ps` and $0.33\ (0.06, 0.61)$ years with `rp`. I conclude that at $t^* = 3$ years, treatment with penicillin extends the restricted mean time to recurrence of cellulitis by about four months.

5 Discussion

I have described two methods for performing the combined test that are both implemented by `stctest`, pseudovalues (`ps`), and RP models (`rp`). The two approaches give similar but not identical results. Which method would I recommend for trial design and analysis?

In essentially all relevant trials, the primary test of the null hypothesis is a Cox or log-rank test of the treatment covariate, with no other covariates in the model. Although in practice a trial may have been designed with stratification on known prognostic and structural factors, such factors are not normally accounted for in the power and sample-size calculations and arguably should not be included in the primary analysis. The issue of covariate adjustment is somewhat controversial, and this is not the place to pursue it.

Accordingly, I recommend the `ps` method when covariate adjustment is not envisioned. The reasons are as follows: First, `ps` does not require the user to choose a suitable "model" from which to estimate ΔRMST, because the pseudovalues method is based on "nonparametric" Kaplan–Meier curves. The `rp` method assumes a particular formulation of the spline model (by default, as already mentioned, `df(5)` and `dftvc(5)`) to estimate the survival curves in each treatment group and hence ΔRMST. The user can alter the spline model via the `df()` and `dftvc()` options of `stctest` if desired. However, I discourage such modifications because they are potentially data driven, which is undesirable in a trial context. I believe the default settings are sufficiently flexible to estimate ΔRMST reliably in the vast majority of trials. Second, the `ps` method is considerably faster to execute than `rp`. Such efficiency is helpful when using the `nperm()` option to check the p-value of the nonstochastic combined test.

If covariate adjustment is deemed essential, I recommend the `rp` method because adjustment for covariates with the `ps` method is done differently for the two components of the combined test, namely, the Cox and RMST models. The Cox test makes a PH assumption, whereas adjusted RMST estimation involves linear regression of the pseudovalues on the covariates and treatment. This does not seem a coherent approach. With the `rp` method, all covariates except treatment are adjusted for in a PH model, with the treatment effect permitted to have non-PH. Further elaboration of this rather complex issue is beyond the scope of this article.

A reviewer pointed out that there are rather few events (15 to be exact) during the third year and subsequent few months of follow-up. If one truncates follow-up at two years, the Cox test of the treatment effect is significant ($p = 0.0071$), with no evidence of non-PH ($p = 0.53$). The combined test gives $\widetilde{p}_{CT} = 0.0105$. This confirms that there is a real treatment effect. Most of the evidence for non-PH appears to arise from the characteristics of the event and censoring times during the third year. However, one would never present an analysis of the data truncated at two years as a primary assessment of the treatment effect, because such an analysis would certainly not have been prespecified in the trial protocol.

For practical use in trial design, Royston and Parmar (2016) suggested a simple, rough-and-ready way to power a trial under PH when the primary test of the null hypothesis is the combined test. A more precise approach to sample-size calculation requires simulation. Work on new commands implementing power and sample-size calculations is under way and will be reported in the *Stata Journal* in due course.

6 Acknowledgments

I thank Professors Kim Thomas and Hywel Williams together with the PATCH1 trial team for permission to use the trial dataset as an example, and I thank Professor Mahesh Parmar and Dr. Tim Morris for helpful comments on the manuscript. I am grateful to a reviewer whose comments prompted me to improve the presentation and the terminology.

7 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.

Cronin, A., L. Tian, and H. Uno. 2016. strmst2 and strmst2pw: New commands to compare survival curves using the restricted mean survival time. *Stata Journal* 16: 702–716.

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. 2015. Estimating the treatment effect in a clinical trial using difference in restricted mean survival time. *Stata Journal* 15: 1098–1117.

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.

———. 2016. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Medical Research Methodology* 16: 16.

Thomas, K. S., A. M. Crook, A. J. Nunn, K. A. Foster, J. M. Mason, J. R. Chalmers, I. S. Nasr, R. J. Brindle, J. English, S. K. Meredith, N. J. Reynolds, D. de Berker, P. S. Mortimer, and H. C. Williams. 2013. Penicillin to prevent recurrent leg cellulitis. *New England Journal of Medicine* 368: 1695–1703.

Trinquart, L., J. Jacot, S. C. Conner, and R. Porcher. 2016. Comparison of treatment effects measured by the hazard ratio and by the ratio of restricted mean survival times in oncology randomized controlled trials. *Journal of Clinical Oncology* 34: 1813–1819.

About the author

Patrick Royston is a medical statistician with 40 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 and tests of treatment effects in trials with a time-to-event outcome, on parametric modeling of survival data, and on novel clinical trial designs.