strbee: Randomization-based efficacy estimator

Ian R. White
MRC Biostatistics Unit, Cambridge, UK
ian.white@mrc-bsu.cam.ac.uk

Sarah Walker
MRC Clinical Trials Unit, London, UK

Abdel Babiker
MRC Clinical Trials Unit, London, UK
abdel.babiker@ctu.mrc.ac.uk

Abstract. strbee analyzes a two-group clinical trial with a survival outcome, in which some subjects may "crossover" to receive the treatment of the other arm. Adjustment for treatment crossover is done by a randomization-respecting method that preserves the intention-to-treat p-value.

Keywords: st0012, clinical trials, treatment changes, randomization–respecting

1 Description

strbee analyzes a two-group clinical trial with a survival outcome, in which some subjects may "crossover" to receive the treatment of the other arm. The method is due to Robins and Tsiatis (1991), and has been developed by White et al. (1997, 1999). The observed event time T is related to an underlying event time U that would have been observed in the absence of treatment through an accelerated life model. The parameter ψ of this model represents the factor by which life is accelerated by treatment and is estimated as the value at which U is balanced between the treatment groups (on a user-specified test).

Estimation is performed by searching over a range of values of ψ and computing the test statistic $Z(\psi)$. Censoring that is noninformative on the T scale becomes informative on the U scale. To cope with this, strbee performs recensoring using potential censoring times.

2 Syntax

strbee *treatvar* [if *exp*] [in *range*] [, test(logrank | wilcoxon | cox |
  weibull | exponential) xo0(*timevar eventvar*) xo1(*timevar eventvar*)
  endstudy(*varname*) savedta(*filename*[, append | replace]) psimin(#)
  psimax(#) psistep(#) tol(#) noci trace list graph level(#)
  graph_options]

strbee using *filename* [, list graph level(#) *graph_options*]

`strbee` is an `st` command; data must be `stset` before running `strbee`. Each subject must have a single record starting at time 0. `treatvar` is the treatment arm, which must have values 0 and 1. Estimation is performed by searching over a range of values of ψ and computing the test statistic $Z(\psi)$.

Censored data present an extra problem for `strbee`, since noninformative censoring of T implies informative censoring of U. To avoid bias, users must specify the potential censoring time for all subjects: `strbee` then computes a recensoring time, which may be earlier than the actual censoring time, and recensors the data.

The first syntax above is the standard syntax. Results for ψ and $Z(\psi)$ may be stored or appended to an existing results file. The second syntax works from previously stored results for ψ and $Z(\psi)$.

3 Options

3.1 Specifying the model

`test(logrank | wilcoxon | cox | weibull | exponential)` specifies the test used to compare U in the two randomized arms. `logrank` and `wilcoxon` are the standard rank tests. The default is the logrank test. `cox`, `weibull`, and `exponential` use the Wald tests from `stcox`, `streg, dist(wei)`, and `streg, dist(exp)`, respectively.

`xo0(`*timevar eventvar*`)` specifies time to crossover and a crossover indicator in arm 0. If not specified, it is assumed that there is no crossover from arm 0 to arm 1.

`xo1(`*timevar eventvar*`)` does the same for arm 1. If not specified, it is assumed that there is no crossover from arm 1 to arm 0.

`endstudy(`*varname*`)` specifies the time of the end of study (the potential censoring time). U values are then recensored at the minimum possible potential censoring time on the U scale, where the minimization is carried out over all treatment profiles possible for the subject's randomized arm.

Note that the potential censoring time must be specified for both censored and uncensored subjects, and that censoring due to random events (e.g., competing risk or loss to follow-up) should ideally be treated differently from censoring at the end of the study. However, if small amounts of random censoring are present, a reasonable approximation is to set the potential censoring time for subjects who are randomly censored equal to the actual time of random censoring.

3.2 Specifying the estimation procedure

`psimin(#)` and `psimax(#)` specify extreme values of the parameter ψ. The defaults are -1 and 1. With the second syntax, they only affect the values plotted on the graph.

psistep(#) with #>0 specifies the step size between the extreme values for a grid search.

psistep(0) specifies an interval-bisection approach for ψ. $Z(\psi)$ is evaluated at psimin() and psimax(). If Z(psimin)< 0 or Z(psimax)> 0, ψ cannot be estimated; a wider range of ψ is probably required. Otherwise, $Z(\psi)$ is evaluated at (psimin+psimax)/2, and ψ is narrowed down to the appropriate half of the interval. After convergence, the procedure is repeated to find the upper and then the lower confidence limit.

Interval bisection may give wrong answers if the test statistic is nondecreasing in ψ, and should always be checked using the graph option.

tol(#) specifies the convergence criterion for interval-bisection estimation. strbee searches until bounds for each solution differ by less than $10^{-\text{tol}}$. tol() also defines the number of decimal places reported. The default is tol(3).

noci suppresses searching for the confidence limits. strbee still reports what limited information it has about them.

level(#) specifies the confidence interval, in percent, for confidence intervals. The default is level(95) or as set by set level; see [U] **23.5 Specifying the width of confidence intervals**.

3.3   Controlling storage of results

savedta(*filename*[, append | replace]) directs the values of ψ and the test statistic to the specified file.

If append is specified, strbee checks that the current test statistics and the stored test statistics were computed using the same model. In this case, the results reported only use test statistics computed in this run of strbee: to get full results, run strbee using *filename*.

3.4   Controlling output

trace gives details on recensoring and prints ψ and the test statistic at each step. It also saves a file rbeetrace.dta containing variables u, du, z0, dz0, z1, dz1, and recens evaluated at the last value of ψ used. Note that with grid search this is the value of psimax() and that with interval-bisection search it is the upper confidence limit without noci or the point estimate with noci.

list lists the values of ψ and the test statistic.

graph graphs the test statistic against ψ. All graph options are allowed.

graph_options are most of the options allowed with graph, twoway; see [G] **graph options**.

4 Example

We will illustrate `strbee` using simulated data based on a randomized controlled trial; see Concorde Coordinating Committee (1994). Patients with asymptomatic HIV infection were randomized to immediate treatment with zidovudine or to deferred treatment with zidovudine. The outcome is time-to-disease progression or death.

An intention-to-treat analysis estimates the effect of applying the policy of immediate zidovudine compared with deferred zidovudine. However, some patients in the deferred arm started zidovudine before symptoms of HIV on the basis of low CD4 cell counts (a prognostic marker for disease progression). Therefore, the intention-to-treat estimate of treatment effect is biased towards the null as an estimate of the efficacy of zidovudine.

4.1 Data

The data are stored in the file `immdef.dta`.

```
. use immdef, clear
(Example data for strbee)

. summarize
     Variable |       Obs        Mean    Std. Dev.        Min        Max

           id |      1000       500.5    288.8194          1       1000
          def |      1000          .5    .5002502          0          1
          imm |      1000          .5    .5002502          0          1
       censyrs|      1000      2.2452    .4529483        1.5          3
           xo |      1000        .189    .3917045          0          1
        xoyrs |      1000    .7770897    .9340944          0          3
         prog |      1000        .312    .4635417          0          1
      progyrs |      1000      1.9325    .6633275    .0215206          3
        entry |      1000       .7548    .4529483          0        1.5

. list id imm xo xoyrs prog progyrs in 1/10, noobs nodisplay

     id     imm      xo      xoyrs    prog     progyrs
      1       1       0          0       0           3
      2       0       1   2.652797       0           3
      3       1       0          0       1    1.737838
      4       1       0          0       1    2.166291
      5       0       1     2.1221       1    2.884646
      6       0       1    .557392       0           3
      7       0       0    2.18947       1     2.18947
      8       1       0          0       1    .9226239
      9       1       0          0       0           3
     10       1       0          0       0           3
```

For example, subject 5 in the deferred arm started treatment ("crossed over") at 2.12 years and progressed at 2.88 years. Subject 2 crossed over at 2.65 years and was censored for progression at 3 years. Subject 3 had crossover censored at time 0 (as did all subjects in the immediate arm) and progressed at 1.74 years.

4.2   Intention-to-treat analysis

First, we estimate the intention-to-treat estimate of the effect of the policy of giving immediate zidovudine from the accelerated failure time model, ignoring treatment changes. An unweighted logrank test using `sts test` gives a borderline significant p-value of 0.056. Using `strbee`, the default method to test for equality of underlying survival times (U) is also the unweighted logrank test (equivalent to the Cox proportional hazards score test).

```
. stset progyrs, failure(prog)
 (output omitted )

. strbee imm
imm=0: 500 subjects, 0 cross-overs
imm=1: 500 subjects, 0 cross-overs
Search method: interval bisection
.............................
Estimating psi from accelerated life model
Method: Robins-Tsiatis method
Test: logrank
Cross-over in arm 0: none
Cross-over in arm 1: none
Treatment variable: imm
End of study variable: none
-------------------------------------------------
               psi       95% confidence interval
-------------------------------------------------
best          -0.147     -0.302       0.002
lower bound   -0.147     -0.302       0.001
upper bound   -0.146     -0.301       0.002
-------------------------------------------------
```

The intention-to-treat ITT estimate is -0.147, meaning that lifetime is used up $\exp(-0.147) = 0.863$ times as fast when on treatment as when off treatment. This has been called an acceleration factor.

Note that the confidence interval spans 0, consistent with the p-value from `sts test`.

4.3   Adjusting for crossovers

The effect of patients in the deferred zidovudine arm starting open label zidovudine is to dilute the estimated efficacy of zidovudine from the intention-to-treat analysis. Estimating the efficacy of zidovudine adjusting for these changes from the control group (`imm = 0`) increases the size of the effect as expected.

(Continued on next page)

```
. strbee imm, xo0(xoyrs xo)
imm=0: 500 subjects, 189 cross-overs
imm=1: 500 subjects, 0 cross-overs
Search method: interval bisection
.............................
Estimating psi from accelerated life model
Method: Robins-Tsiatis method
Test: logrank
Cross-over in arm 0: xoyrs xo
Cross-over in arm 1: none
Treatment variable: imm
End of study variable: none
-------------------------------------------------
               psi      95% confidence interval
-------------------------------------------------
best          -0.185    -0.367       0.004
lower bound   -0.186    -0.367       0.004
upper bound   -0.185    -0.366       0.005
-------------------------------------------------
```

The point estimate is now -0.185, further from the null than the intention-to-treat estimate, but the upper confidence limit remains just above 0, reflecting the fact that the p-value is unchanged by this procedure.

In some situations, we might want to adjust for treatment crossovers from both arms, which can be specified separately using xo0() and xo1().

4.4   Adjusting for censoring

The trial lasted 3 years, with staggered entry over the first 1.5 years. Variable censyrs gives the time from entry to the end of the 3 years.

So far, censoring indicators on the progression time scale in the trial (T) scale have been passed over to the underlying failure time (U) scale. Unfortunately, uninformative censoring on the T scale may be informative on the underlying U time scale, because patients receiving the superior treatment have their underlying lifetime U extended so that they are censored, whereas patients with the same underlying failure time U receiving the inferior treatment may have the event observed. Thus, on the U scale, censoring is informative with respect to the treatment group. One solution is to "recensor" the underlying failure times, U, by the minimum underlying failure time that could have been observed for each patient across all possible history of treatment changes. This requires passing to strbee the potential censoring time for each patient, which is the time from trial entry to the planned end of study. Recensoring is therefore associated with a loss of information, but a smaller bias in estimating the treatment effect.

For our data, the trial lasted 3 years, with staggered entry over the first 1.5 years. The potential censoring time is variable censyrs, which ranges from 3 years for the first patients to 1.5 years for the last patients.

```
. l id imm censyrs xo xoyrs prog progyrs in 1/10, noobs nodis
     id      imm    censyrs       xo       xoyrs     prog      progyrs
      1        1          3        0           0        0            3
      2        0          3        1    2.652797        0            3
      3        1          3        0           0        1     1.737838
      4        1          3        0           0        1     2.166291
      5        0          3        1      2.1221        1     2.884646
      6        0          3        1     .557392        0            3
      7        0          3        0     2.18947        1      2.18947
      8        1          3        0           0        1     .9226239
      9        1          3        0           0        0            3
     10        1          3        0           0        0            3
```

The result of strbee with recensoring is shown below. In this case, the point estimate is slightly different. We also used the savedta() option which stores the results for future use.

```
. strbee imm, xo0(xoyrs xo) endstudy(censyrs) savedta(recens)
imm=0: 500 subjects, 189 cross-overs
imm=1: 500 subjects, 0 cross-overs
Search method: interval bisection
.............................
Estimating psi from accelerated life model
Method: Robins-Tsiatis method
Test: logrank
Cross-over in arm 0: xoyrs xo
Cross-over in arm 1: none
Treatment variable: imm
End of study variable: censyrs
------------------------------------------------
               psi       95% confidence interval
------------------------------------------------
best          -0.181     -0.350         0.002
lower bound   -0.182     -0.351         0.002
upper bound   -0.181     -0.350         0.003
------------------------------------------------

file recens.dta saved
```

4.5   Advanced use and checking for multiple crossovers

In practice, we recommend exploring the graph of $Z(\psi)$ against ψ in greater detail, especially to check that there really is a unique solution to $Z(\psi)=0$. In doing this, we will illustrate some more advanced features of strbee.

We will explore the range -0.5 to 0.1, which contains the 95% confidence interval reported above. We will use a grid search, evaluating $Z(\psi)$ at values of ψ between -0.5 and 0.1 in steps of 0.02, and draw a graph. We will append the results to those saved above.

```
. strbee imm, xo0(xoyrs xo) endstudy(censyrs) savedta(recens,append) /*
>    */ psimin(-0.5) psimax(0.1) psistep(0.02) graph /*
>    */ title(RBEE with recensoring)
imm=0: 500 subjects, 189 cross-overs
imm=1: 500 subjects, 0 cross-overs
Search method: grid search
.............................
Estimating psi from accelerated life model
Method: Robins-Tsiatis method
Test: logrank
Cross-over in arm 0: xoyrs xo
Cross-over in arm 1: none
Treatment variable: imm
End of study variable: censyrs
------------------------------------------------
              psi       95% confidence interval
------------------------------------------------
best         -0.184    -0.351       0.003
lower bound  -0.200    -0.360       0.000
upper bound  -0.180    -0.340       0.020
------------------------------------------------
Appending results from recens.dta

2 duplicate records deleted
file recens.dta saved
```

The point estimate is different from before only because of the coarse grid used; ψ is really only estimated as lying between -0.2 and -0.18. We can get the best results from the combined analysis:

```
. strbee using recens, graph psimin(-0.5) psimax(0.1)
(Example data for strbee)

Estimating psi from accelerated life model
Method: Robins-Tsiatis method
Test: logrank
Cross-over in arm 0: xoyrs xo
Cross-over in arm 1: none
Treatment variable: imm
End of study variable: censyrs
------------------------------------------------
              psi       95% confidence interval
------------------------------------------------
best         -0.181    -0.350       0.002
lower bound  -0.182    -0.351       0.002
upper bound  -0.181    -0.350       0.003
------------------------------------------------
```

The point estimate is the same as before. Of greater interest is the graph of $Z(\psi)$ against ψ, which shows no evidence of nonmonotonicity.
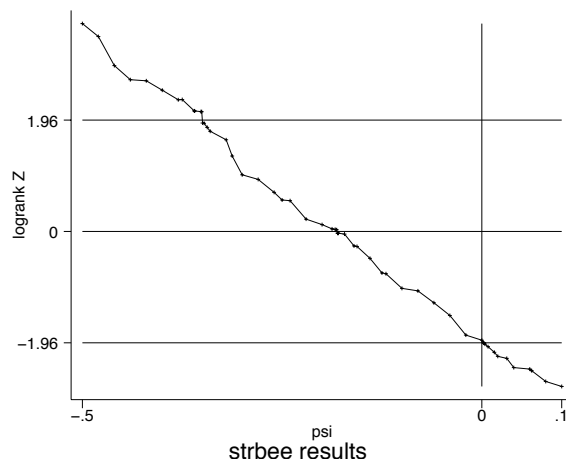
Figure 1: Graph of $Z(\psi)$ versus ψ for the example data.

5 Methods and formulas

Let the observed event time for patient i be $T_i = T_{0i} + T_{1i}$, where T_{0i} and T_{1i} are the lengths of time that the patient spent off and on treatment before the event, respectively. For example, in arm 0, T_{0i} equals the time to crossover, or the time to event if there was no crossover before the event. If time to crossover is censored before the event time, strbee assumes that no crossover occurred before the event.

The observed event time T_i is related to treatment-free event time U_i by a causal model

$$U_i = T_{0i} + e^{\psi_0} T_{1i}$$

where ψ_0 is the true causal parameter. The basic assumption is that $U \perp\!\!\!\perp R$, where $R_i = 0/1$ indicates the randomized treatment arm.

For any given value of ψ, we test the hypothesis $\psi_0 = \psi$ by computing

$$U_i(\psi) = T_{0i} + e^{\psi} T_{1i}$$

and calculating $Z(\psi)$ as the test statistic for the hypothesis $U(\psi) \perp\!\!\!\perp R$. Specifically, the logrank, Wilcoxon, Cox, exponential, and Weibull tests are performed by

```
. sts test treatvar
. sts test treatvar, wilcoxon
. stcox treatvar
. streg treatvar, dist(exponential) time
. streg treatvar, dist(weibull) time
```

With the parametric tests (exponential and Weibull), the point estimate is the value of ψ for which $Z(\psi) = 0$. With the nonparametric tests (logrank, Wilcoxon, and Cox),

$Z(\psi)$ is a step function, and the point estimate is the value of ψ for which $Z(\psi)$ crosses 0. In all cases, the $100(1-\alpha)\%$ confidence interval is the set $\left\{\psi : |Z(\psi)| < z_{1-\alpha/2}\right\}$.

Let the potential censoring time for individual i be C_i, known for uncensored as well as censored individuals. To avoid bias from informative censoring of $U_i(\psi)$, $U_i(\psi)$ is recensored at its earliest possible censoring time

$$D_i^*(\psi) = \min\left(C_i, e^\psi C_i\right)$$

Thus, if $D_i^*(\psi) < U_i(\psi)$, then $U_i(\psi)$ is replaced by $D_i^*(\psi)$ and its censoring indicator is replaced by 0. In any randomized treatment arm in which subjects do not crossover, informative censoring of $U_i(\psi)$ cannot occur, and so recensoring is not applied. White et al. (1999) discuss recensoring in detail.

The method used for the interval-bisection search is as follows. `strbee` first evaluates $Z(\psi)$ at $\psi = $ `psimin` and `psimax`. If they have the same sign, then the procedure stops because there is probably no solution in this range. Otherwise, `strbee` next evaluates $Z(\psi)$ at $\psi = $ `(psimin+psimax)/2`, and it continues to bisect intervals until the interval width is less than $10^{-\text{tol}}$. It then uses the already computed values of $Z(\psi)$ to put bounds on the lower confidence limit, and uses another interval-bisection search between these bounds. Finally, it repeats the same procedure for the upper confidence limit.

A grid search may show that repeated solutions of $Z(\psi) = 0$ exist. To find a single point estimate, `strbee` uses an approach proposed by White et al. (1999). For example, suppose $Z(\psi) < 0$ for $\psi < 0.2$ and for $0.3 < \psi < 0.5$. The total length of the range for which $Z(\psi) < 0$ is the same as if $Z(\psi)$ only crossed at $\psi = 0.4$, so the point estimate is taken as 0.4.

6 Limitations

`strbee` can account only for a single change in treatment to the treatment of the opposite trial arm. For methods for data with multiple treatment changes, see White et al. (1999). For methods for data with changes to nontrial treatments, see White et al. (1997).

7 Saved results

`strbee` saves the following scalars in `r()`:

`r(psi)`	estimate of ψ
`r(psiupp)`	estimate of upper confidence limit around ψ
`r(psilow)`	estimate of lower confidence limit around ψ

8   References

Concorde Coordinating Committee. 1994. Concorde: MRC/ANRS randomised double-blind controlled trial of immediate and deferred zidovudine in symptom-free HIV infection. *Lancet* 343: 871–881.

Robins, J. M. and A. A. Tsiatis. 1991. Correcting for non-compliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics Theory and Methods* 20: 2609–2631.

White, I. R., A. G. Babiker, S. Walker, and J. H. Darbyshire. 1999. Randomisation-based methods for correcting for treatment changes: examples from the Concorde trial. *Statistics in Medicine* 18: 2617–2634.

White, I. R., S. Walker, A. G. Babiker, and J. H. Darbyshire. 1997. Impact of treatment changes on the interpretation of the Concorde trial. *AIDS* 11: 999–1006.

About the Authors

Ian White is a Biostatistician at the Medical Research Council (MRC) Biostatistics Unit in Cambridge, UK.

Sarah Walker is a Biostatistician at the MRC Clinical Trials Unit in London, UK.

Abdel Babiker is head of the HIV division of the MRC Clinical Trials Unit in London, UK.