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	
1-year subscription	$ 79	1-year subscription	$115
2-year subscription	$155	2-year subscription	$225
3-year subscription	$225	3-year subscription	$329
3-year subscription (electronic only)	$210	3-year subscription (electronic only)	$210
1-year student subscription	$ 48	1-year student subscription	$ 79
1-year university library subscription	$ 99	1-year university library subscription	$135
2-year university library subscription	$195	2-year university library subscription	$265
3-year university library subscription	$289	3-year university library subscription	$395
1-year institutional subscription	$225	1-year institutional subscription	$259
2-year institutional subscription	$445	2-year institutional subscription	$510
3-year institutional subscription	$650	3-year institutional subscription	$750

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 (2012)
12, Number 4, pp. 674–687

Simulating complex survival data

Michael J. Crowther
Department of Health Sciences
University of Leicester
Leicester, UK
michael.crowther@le.ac.uk

Paul C. Lambert
Department of Health Sciences
University of Leicester
Leicester, UK
and
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden

Abstract. Simulation studies are essential for understanding and evaluating both current and new statistical models. When simulating survival times, one often assumes an exponential or Weibull distribution for the baseline hazard function, with survival times generated using the method of Bender, Augustin, and Blettner (2005, *Statistics in Medicine* 24: 1713–1723). Assuming a constant or monotonic hazard can be considered too simplistic and can lack biological plausibility in many situations. We describe a new user-written command, `survsim`, which allows the user to simulate survival times from two-component parametric mixture models, providing much more flexibility in the underlying hazard. Standard parametric distributions can also be used, including the exponential, Weibull, and Gompertz. Furthermore, survival times can be simulated from the all-cause distribution of cause-specific hazards for competing risks by using the method of Beyersmann et al. (2009, *Statistics in Medicine* 24: 956–971). A multinomial distribution is used to create the event indicator, whereby the probability of experiencing each event at a simulated time t is the cause-specific hazard divided by the all-cause hazard evaluated at time t. Baseline covariates can be included in all scenarios. We also describe the extension to incorporate nonproportional hazards in standard parametric and competing-risks scenarios.

Keywords: st0275, survsim, simulation, survival analysis, mixture models, competing risks

1 Introduction

Simulation studies are commonly used to evaluate the performance of both current and newly developed statistical models (Burton et al. 2006). Within the survival analysis field, either the exponential distribution, which assumes a constant underlying hazard function, or a Weibull distribution, which assumes a monotonically increasing or de-

creasing hazard, is used and implemented with the methods of Bender, Augustin, and Blettner (2005). These choices can be considered too simplistic and can lack biological plausibility to accurately reflect many real-world datasets. For example, in the analysis of cancer survival data, a turning point is often observed in the hazard function. Furthermore, despite the Cox model (Cox 1972) being the most commonly used method of survival analysis, it is not possible to simulate from a semiparametric model.

A class of finite mixture models has been proposed to increase the flexibility of fully parametric survival models (McLachlan and McGiffin 1994). We describe the `survsim` command, which uses a special case of these models to simulate survival data from two-component parametric mixture distributions, incorporating much more flexibility in the underlying hazard function. Survival times can also be simulated from standard parametric models, including the exponential, Weibull, and Gompertz. Furthermore, survival times can be simulated from the all-cause distribution of cause-specific hazards for competing risks under the method of Beyersmann et al. (2009). A multinomial distribution is used to create the event indicator, whereby the probability of experiencing each event at a simulated time t is the cause-specific hazard divided by the all-cause hazard evaluated at time t. Baseline covariates can be included in all scenarios. We also describe the extension to incorporate nonproportional hazards in standard parametric and competing-risks scenarios.

2 Simulating survival times

2.1 Two-component parametric mixture distribution

We begin by defining the survival function of the two-component Weibull mixture, with $\lambda_1, \lambda_2, \gamma_1, \gamma_2 > 0$ and $0 \leq p \leq 1$

$$S_0(t) = p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2}) \tag{1}$$

where $\{\lambda_1, \lambda_2\}$ and $\{\gamma_1, \gamma_2\}$ are scale and shape parameters, respectively. p represents the mixing parameter. Transforming to the cumulative hazard scale, we get

$$H_0(t) = -\log\{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})\}$$

Differentiating with respect to t, we obtain the baseline hazard function:

$$h_0(t) = \frac{\lambda_1\gamma_1 t^{\gamma_1-1}p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})}$$

Proportional hazards can then be simply incorporated by

$$h(t) = \frac{\lambda_1\gamma_1 t^{\gamma_1-1}p\exp(-\lambda_1 t^{\gamma_1}) + \lambda_2\gamma_2 t^{\gamma_2-1}(1-p)\exp(-\lambda_2 t^{\gamma_2})}{p\exp(-\lambda_1 t^{\gamma_1}) + (1-p)\exp(-\lambda_2 t^{\gamma_2})}\exp(\mathbf{x}\boldsymbol{\beta}) \tag{2}$$

where $\mathbf{x}$ is a vector of covariates and β is the corresponding vector of regression coefficients. Transforming back to the survival scale gives

$$S(t) = S_0(t)^{\exp(\mathbf{x}\boldsymbol{\beta})}$$

With the method of Bender, Augustin, and Blettner (2005), we use the relationship between the survival function and the cumulative distribution function, whereby

$$S(t) = 1 - F(t), \qquad \text{where} \quad F \sim U(0,1) \tag{3}$$

We then make n draws from $F \sim U(0,1)$, substituting each into (3) and solving for t. Under most standard parametric models, we can directly solve for t, but under a mixture model, we must use root-finding techniques such as Newton–Raphson iterations to solve for t and hence generate the survival times. Figure 1 displays a variety of complex hazard functions that can be simulated from the mixture model.

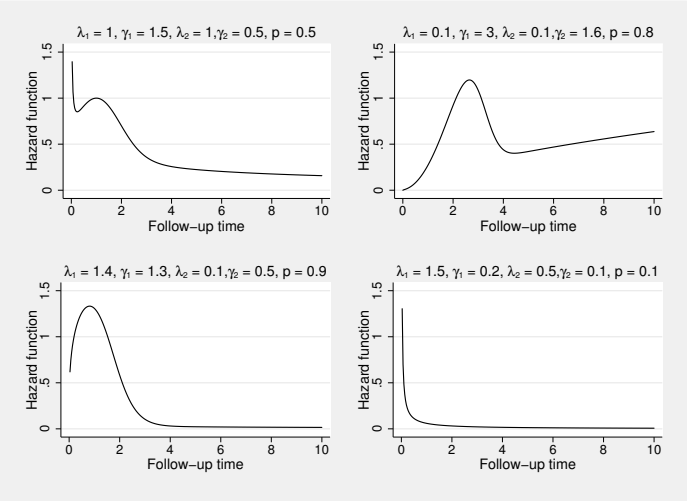

Figure 1. Complex hazard functions that can be simulated from the two-component Weibull mixture

The complex hazard functions displayed in figure 1 will allow the user to better reflect those observed in many real datasets.

2.2 Simulating competing-risks data

The Newton–Raphson approach to simulating survival times can also be applied to the competing-risks setting. For a general introduction to competing risks, we refer the reader to Putter, Fiocco, and Geskus (2007). Under the method of Beyersmann et al. (2009), survival times can be generated from the all-cause distribution of K cause-specific hazard functions. We define the kth cause-specific hazard function to be

$$h_k(t) = h_{0k}(t) \exp\left(\mathbf{x}_k \boldsymbol{\beta}_k\right)$$

whereby each cause-specific baseline hazard $h_{0k}(t)$ can be the exponential, Weibull, or Gompertz parametric distribution. Separate baseline covariates, $\mathbf{x}_k$, can be included in each cause-specific hazard function with corresponding regression coefficients $\boldsymbol{\beta}_k$. The all-cause hazard is therefore

$$h_{\text{all}}(t) = \sum_{k=1}^{K} h_k(t)$$

Once the survival times are generated, a multinomial distribution is used to create the event indicator, whereby the probability of experiencing event k at a simulated time t is the cause-specific hazard of event k divided by the all-cause hazard evaluated at time t.

2.3 Time-dependent effects

It is also desirable to incorporate time-dependent effects when assessing survival methods. Time-dependent effects may occur, for example, when a treatment effect diminishes with time. Incorporating them can be done quite simply under the standard parametric models. Under an exponential or Weibull model, covariates can be interacted with log time. This enables us to use the method of Bender, Augustin, and Blettner (2005) because we can directly solve for t. Similarly, under a Gompertz distribution, covariates can be interacted with time. This can also be implemented for each cause-specific hazard under a competing-risks model. We discuss the extension to incorporating time-dependent effects in the mixture models in section 5.

2.4 Censoring

When undertaking simulation studies, one often assumes that the censoring distribution is uniform or follows an exponential distribution. The same procedures described above can also be used to generate a censoring distribution to better reflect those seen in observed datasets. The censoring indicator can then be constructed from the minimum of a simulated survival time and a simulated censoring time.

3 The survsim command

3.1 Syntax

```
survsim newvar1 [newvar2] [, n(#) lambdas(numlist) gammas(numlist)
   distribution(exponential|weibull|gompertz) covariates(varname # [#
   ...] ...)  tde(varname # [# ...] ...)  mixture pmix(#) cr ncr(#)
   centol(#) showdiff]
```

newvar1 specifies the new variable name to contain the generated survival times. *newvar2* is required when generating competing-risks data to create the status indicator.

3.2 Options

`n(#)` specifies the number of survival times to generate. The default is the number of observations in the current dataset.

`lambdas(`*numlist*`)` defines the scale parameters in the exponential, Weibull, and Gompertz distributions. The number of values required depends on the model choice. The default is one number corresponding to a standard parametric distribution. Under a `mixture` model, two values are required. Under a competing-risks (`cr`) model, the number of values is defined by `ncr()`.

`gammas(`*numlist*`)` defines the shape parameters of the Weibull or Gompertz parametric distributions. The number of entries must be equal to that of `lambdas()`.

`distribution(exponential | weibull | gompertz)` specifies the parametric survival distribution to use. The default is `distribution(weibull)`.

`covariates(`*varname* # [# ...] ...`)` defines baseline covariates to be included in the linear predictor of the survival model, along with the value of the corresponding coefficient. For example, a treatment variable coded 0/1 can be included, with a log hazard-ratio of 0.5, by `covariates(treat 0.5)`. The variable `treat` must be in the dataset before `survsim` is run. If `cr` is used with `ncr(4)`, then a value for each covariate must be input for each competing risk, for example, `covariates(treat 0.5 -0.2 0.1 0.25)`.

`tde(`*varname* # [# ...] ...`)` creates nonproportional hazards by interacting covariates with log time under a Weibull or exponential model or with time under a Gompertz model. This option is not available under a mixture model. Values should be entered, for example, as `tde(trt 0.5)`.

`mixture` specifies that survival times be simulated from a two-component mixture model. `lambdas()` and `gammas()` must be of length 2.

`pmix(#)` defines the value of the mixture parameter. The default is `pmix(0.5)`.

`cr` specifies that survival times be simulated from the all-cause distribution of `ncr()` cause-specific hazards.

`ncr(#)` defines the number of competing risks, that is, the number of cause-specific hazards. `lambdas()` and `gammas()` must be of the length defined by `ncr()`.

`centol(#)` specifies the tolerance of the Newton–Raphson scheme. The default is `centol(0.0001)`.

`showdiff` shows the maximum difference in estimates between iterations of the Newton–Raphson scheme. This can be used to monitor convergence.

4 Example use of survsim

We illustrate the use of `survsim` through some simple simulation studies.

4.1 Standard parametric survival model

The first example illustrates a somewhat standard simulation study. We simulate survival times from a baseline Weibull distribution with an increasing hazard function. We can incorporate a constant treatment effect by first generating a binary treatment group indicator and defining a log hazard-ratio of -0.5, that is, a hazard ratio of 0.607, indicating a beneficial treatment effect reducing the event rate by 39.3%. We set the seed for reproducibility and conduct 1,000 replicates, analyzing bias and coverage of the treatment effect estimate. We apply a maximum follow-up time of five years.

```
. set seed 6765327
. program simstudy1, rclass
  1.         clear
  2.         set obs 1000
  3.         generate trt = rbinomial(1,0.5)
  4.         survsim stime, distribution(weibull) lambdas(0.1) gammas(1.5)
> covariates(trt -0.5)
  5.         generate died = stime <= 5
  6.         replace stime = 5 if died == 0
  7.         stset stime, failure(died = 1)
  8.         streg trt, distribution(weibull) nohr
  9.         return scalar loghr = _b[trt]
 10.         return scalar seloghr = _se[trt]
 11. end
. simulate loghr = r(loghr) seloghr = r(seloghr), reps(1000) nodots nolegend:
> simstudy1
. /* Bias */
. generate bias = loghr - (-0.5)
. summarize bias, meanonly
. display r(mean)
-.00340769
. /* Coverage */
. generate cov = (loghr + invnorm(0.975)*seloghr>-0.5 & loghr -
> invnorm(0.975)*seloghr<-0.5)
. tabulate cov
```

cov	Freq.	Percent	Cum.
0	50	5.00	5.00
1	950	95.00	100.00
Total	1,000	100.00	

We observe a minimal bias in the estimates of the log hazard-ratio of -0.003 and a coverage of 95%.

4.2 Two-component parametric survival model

Now consider the German breast cancer dataset available by typing `webuse brcancer`. This dataset consists of 686 patients, randomized with 246 to receive hormonal therapy and 440 to receive a placebo. The outcome is recurrence-free survival, of which 299 patients experienced the event of interest. We first fit a Weibull survival model investigating the effect of hormonal therapy, and we obtain the fitted survival function. We then compare this with a two-component Weibull–Weibull mixture model [described in equations (1) to (2)] by using the `stmix` command available from the Statistical Software Components archive (Crowther and Lambert 2011). Figure 2 displays the fitted values from both models.

```
. webuse brcancer, clear
(German breast cancer data)
. stset rectime, failure(censrec) scale(365.25)
     failure event:  censrec != 0 & censrec < .
obs. time interval:  (0, rectime]
 exit on or before:  failure
    t for analysis:  time/365.25
------------------------------------------------------------------------------
      686  total obs.
        0  exclusions
------------------------------------------------------------------------------
      686  obs. remaining, representing
      299  failures in single record/single failure data
 2111.978  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  7.279945
. sts generate kmsurv = s, by(hormon)
. streg hormon, distribution(weibull) nolog noheader
        failure _d:  censrec
  analysis time _t:  rectime/365.25
```

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hormon	.6748666	.0842414	-3.15	0.002	.528403	.8619273
_cons	.1113439	.0121783	-20.07	0.000	.0898599	.1379644
/ln_p	.250997	.0496958	5.05	0.000	.1535949	.348399
p	1.285306	.0638744			1.166018	1.416798
1/p	.7780247	.0386646			.7058172	.8576193

```
. predict surv_w, surv
```

```
. stmix hormon, distribution(weibweib) nolog
Obtaining initial values:
Fitting full model:
Mixture Weibull-Weibull proportional hazards regression
Log likelihood = -843.05585                     Number of obs   =        686
```

	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
hormon	.6917148	.0864138	-2.95	0.003	.5414883	.8836192
_cons	1	(omitted)				
logit_p_mix						
_cons	.9787895	.290526	3.37	0.001	.4093689	1.54821
ln_lambda1						
_cons	-3.721406	.7223363	-5.15	0.000	-5.13716	-2.305653
ln_gamma1						
_cons	.626371	.189829	3.30	0.001	.2543129	.9984291
ln_lambda2						
_cons	-1.14564	.1566929	-7.31	0.000	-1.452753	-.838528
ln_gamma2						
_cons	.9187288	.1159475	7.92	0.000	.6914759	1.145982

```
. predict surv_ww, survival
```

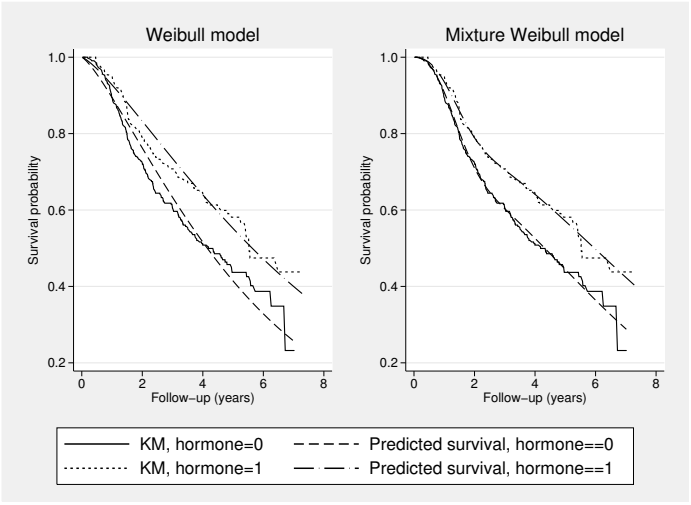

Figure 2. Comparison of predicted survival curves

It is evident from figure 2 that the two-component Weibull mixture model provides a better fit to the observed data. However, although they are useful indicators of im-

proved fit, plots of fitted values overlaid on Kaplan–Meier curves should be interpreted with caution. Quite often, we observe sparse data in the tails, and in such scenarios, fitted values should not be overinterpreted. The majority of the data in this example occurs in the first four years of follow-up, where the fitted values appear to fit well in the two-component mixture model as compared with the Weibull model. This motivating example illustrates a setting where a more complex underlying hazard function is more biologically plausible than a standard parametric model. We can investigate the performance of the two-component mixture model through simulation by using parameter estimates obtained from the `stmix` output. For comparison, we also fit a standard Weibull survival model.

```
. local loghr = [xb][hormon]
. local l1 = exp([ln_lambda1][_cons])
. local g1 = exp([ln_gamma1][_cons])
. local l2 = exp([ln_lambda2][_cons])
. local g2 = exp([ln_gamma2][_cons])
. local pmix = invlogit([logit_p_mix][_cons])
. set seed 878764
. program simstudy2, rclass
  1.         syntax [ , PMIX(real 0.5) L1(real 0.1) L2(real 0.1) G1(real 1)
> G2(real 1) loghr(real 1)]
  2.         clear
  3.         set obs 1000
  4.         generate trt = rbinomial(1,0.5)
  5.         survsim stime, mixture distribution(weibull) lambdas(`l1´ `l2´)
> gammas(`g1´ `g2´) pmix(`pmix´) covariates(trt `loghr´)
  6.         generate died = stime <= 5
  7.         replace stime = 5 if died == 0
  8.         stset stime, failure(died = 1)
  9.         streg trt, distribution(weibull) nohr
 10.         return scalar trt_w = _b[trt]
 11.         return scalar setrt_w = _se[trt]
 12.         cap constraint drop _all
 13.         stmix trt, distribution(weibweib)
 14.         return scalar trt_ww = [xb][trt]
 15.         return scalar setrt_ww = [xb]_se[trt]
 16. end
. simulate trt_w = r(trt_w) setrt_w = r(setrt_w) trt_ww = r(trt_ww)
> setrt_ww = r(setrt_ww), reps(500): simstudy2, pmix(`pmix´) l1(`l1´) l2(`l2´)
> g1(`g1´) g2(`g2´) loghr(`loghr´)
  (output omitted)
. /* Bias */
. generate bias_trt_w = trt_w - (`loghr´)
. generate bias_trt_ww = trt_ww - (`loghr´)
. summarize bias*
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  bias_trt_w |       500   -.0121194    .0931462   -.255441   .2476663
 bias_trt_ww |       500   -.0023753    .0908162   -.242304   .2496307
```

We observe very small levels of bias under both the Weibull and the two-component mixture model of -0.012 and -0.002, respectively.

4.3 Time-dependent effects

We now illustrate the incorporation of a time-dependent effect. This can be done by using a standard Weibull survival model that illustrates a diminishing treatment effect, as follows. The true model is defined as

$$h(t) = \lambda\gamma t^{\gamma-1} \exp\{\beta X_i + \phi X_i \times \log(t)\} \tag{4}$$

We simulate one dataset and fit a flexible parametric survival model (see Royston and Parmar [2002] and Royston and Lambert [2011]), allowing for a time-dependent hazard ratio for the effect of treatment. Flexible parametric models are fit on the log cumulative-hazard scale by using restricted cubic splines. Figure 3 displays the predicted time-dependent hazard ratio. For comparison, we also show the estimate of the time-independent hazard ratio.

```
. set seed 6765327
. clear
. set obs 10000
obs was 0, now 10000
. generate trt = rbinomial(1,0.5)
. survsim stime, distribution(weibull) lambdas(0.1) gammas(1.5)
> covariates(trt -0.5) tde(trt 0.15)
. generate died = stime <= 5
. replace stime = 5 if died == 0
(3869 real changes made)
. stset stime, failure(died = 1)
     failure event:  died == 1
obs. time interval:  (0, stime]
 exit on or before:  failure
--------------------------------------------------------------------------------
    10000  total obs.
        0  exclusions
--------------------------------------------------------------------------------
    10000  obs. remaining, representing
     6131  failures in single record/single failure data
 35805.85  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =         5
. stpm2 trt, scale(h) df(3) tvc(trt) dftvc(1) nolog
Log likelihood = -11350.047                     Number of obs   =      10000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
trt	-.4418441	.0287258	-15.38	0.000	-.4981456	-.3855427
_rcs1	1.031265	.0173282	59.51	0.000	.9973024	1.065228
_rcs2	-.0162161	.0133391	-1.22	0.224	-.0423603	.0099282
_rcs3	-.0024802	.0054753	-0.45	0.651	-.0132117	.0082513
_rcs_trt1	.0866731	.0256407	3.38	0.001	.0364183	.1369279
_cons	-.6103966	.0183731	-33.22	0.000	-.6464072	-.5743859

```
. predict hr, hrnumer(trt 1) ci
```

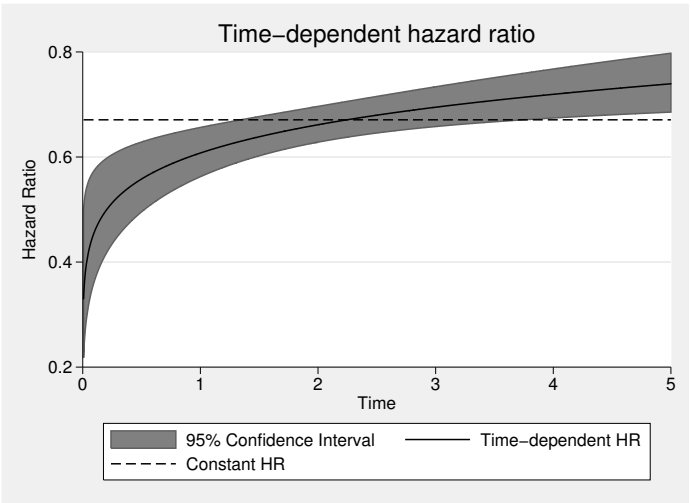

Figure 3. Time-dependent hazard ratio

In this example, we have simulated a time-dependent effect on the hazard scale and have fit a model estimating a time-dependent effect on the cumulative hazard scale. This simulation study could be extended to investigate any biases that arise when the modeling scale differs from that of the data-generating process.

4.4 Competing risks

Finally, we demonstrate the simulation of competing-risks data under a cause-specific hazards model. We define two cause-specific hazards using Weibull-distributed baseline hazard functions, the first with an increasing hazard and the second with a decreasing one. We specify a competing-risks setting by using the `cr` and `ncr()` options. Cause-specific Weibull survival models are then fit to illustrate the method. Censoring is applied after 15 years. Using `stcompet` (Coviello and Boggess 2004), we show the predicted cumulative incidence function for each cause in figure 4.

```
. set seed 6765327
. clear
. set obs 10000
obs was 0, now 10000
. generate trt = rbinomial(1,0.5)
. survsim stime event, distribution(weibull) cr ncr(2) lambdas(0.1 0.1)
> gammas(1.5 0.5) covariates(trt -0.5 0.5)
. replace event = 0 if stime>15
(78 real changes made)
. stset stime, failure(event==1)
     failure event:  event == 1
obs. time interval:  (0, stime]
 exit on or before:  failure
  (output omitted)
. streg trt, distribution(weibull) nohr nolog noheader
        failure _d:  event == 1
  analysis time _t:  stime
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
trt	-.5146497	.0234288	-21.97	0.000	-.5605694	-.46873
_cons	-2.25095	.0274368	-82.04	0.000	-2.304725	-2.197175
/ln_p	.3841366	.0087728	43.79	0.000	.3669421	.401331
p	1.468346	.0128815			1.443314	1.493812
1/p	.6810384	.0059746			.6694285	.6928497

```
. stcompet ci1 = ci, compet1(2) by(trt)
. stset stime, failure(event==2)
     failure event:  event == 2
obs. time interval:  (0, stime]
 exit on or before:  failure
  (output omitted)
. streg trt, distribution(weibull) nohr nolog noheader
        failure _d:  event == 2
  analysis time _t:  stime
```

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
trt	.4793291	.0424895	11.28	0.000	.3960512	.562607
_cons	-2.270358	.0357084	-63.58	0.000	-2.340345	-2.200371
/ln_p	-.7248368	.0188308	-38.49	0.000	-.7617446	-.6879291
p	.4844036	.0091217			.4668513	.5026159
1/p	2.064394	.0388742			1.989591	2.14201

```
. stcompet ci2 = ci, compet1(1) by(trt)
```

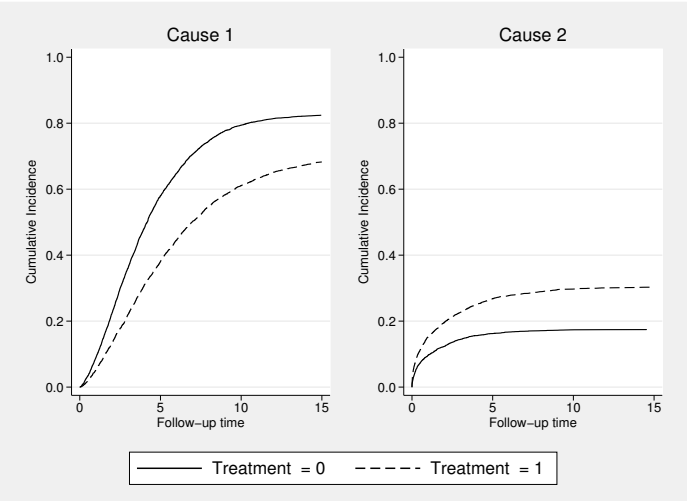

Figure 4. Cumulative incidence

Figure 4 displays the beneficial treatment effect simulated for cause 1 and the detrimental treatment effect simulated for cause 2. The number of competing risks is not limited in `survsim`, and each cause-specific hazard can be extended to include time-dependent effects.

5 Conclusion

We described a flexible tool to simulate a variety of complex survival data. We hope it will be useful not only to generate more realistic and biologically plausible survival data but also to better assess statistical models and improve understanding of the data-generating processes underlying survival models.

Extension to incorporate time-dependent effects within the mixture model framework requires numerical integration to evaluate the cumulative hazard. This draws parallels with the simulation of joint longitudinal and survival data, which is currently under development. Future work also includes the addition of cure proportions and frailty distributions.

6 Acknowledgment

Michael Crowther was funded by a National Institute for Health Research methodology fellowship (RP-PG-0407-10314).

7 References

Bender, R., T. Augustin, and M. Blettner. 2005. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 24: 1713–1723.

Beyersmann, J., A. Latouche, A. Buchholz, and M. Schumacher. 2009. Simulating competing risks data in survival analysis. *Statistics in Medicine* 28: 956–971.

Burton, A., D. G. Altman, P. Royston, and R. L. Holder. 2006. The design of simulation studies in medical statistics. *Statistics in Medicine* 25: 4279–4292.

Coviello, V., and M. Boggess. 2004. Cumulative incidence estimation in the presence of competing risks. *Stata Journal* 4: 103–112.

Cox, D. R. 1972. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* 34: 187–220.

Crowther, M. J., and P. Lambert. 2011. stmix: Stata module to fit two-component parametric mixture survival models. Statistical Software Components S457339, Department of Economics, Boston College.
http://ideas.repec.org/c/boc/bocode/s457339.html.

McLachlan, G. J., and D. C. McGiffin. 1994. On the role of finite mixture models in survival analysis. *Statistics Methods in Medical Research* 3: 211–226.

Putter, H., M. Fiocco, and R. B. Geskus. 2007. Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine* 26: 2389–2430.

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.

About the authors

Michael Crowther is a research assistant in medical statistics and a part-time PhD student. His main research interest is the joint modeling of longitudinal and survival data.

Paul Lambert is a reader in medical statistics. His main interest is in the development and application of methods in population-based cancer research.