



The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

THE STATA JOURNAL

Editors

H. JOSEPH NEWTON
Department of Statistics
Texas A&M University
College Station, Texas
editors@stata-journal.com

NICHOLAS J. COX
Department of Geography
Durham University
Durham, UK
editors@stata-journal.com

Associate Editors

CHRISTOPHER F. BAUM, Boston College
NATHANIEL BECK, New York University
RINO BELLOCCO, Karolinska Institutet, Sweden, and
University of Milano-Bicocca, Italy
MAARTEN L. BUIS, University of Konstanz, Germany
A. COLIN CAMERON, University of California–Davis
MARIO A. CLEVES, University of Arkansas for
Medical Sciences
WILLIAM D. DUPONT, Vanderbilt University
PHILIP ENDER, University of California–Los Angeles
DAVID EPSTEIN, Columbia University
ALLAN GREGORY, Queen's University
JAMES HARDIN, University of South Carolina
BEN JANN, University of Bern, Switzerland
STEPHEN JENKINS, London School of Economics and
Political Science
ULRICH KOHLER, University of Potsdam, Germany

FRAUKE KREUTER, Univ. of Maryland–College Park
PETER A. LACHENBRUCH, Oregon State University
JENS LAURITSEN, Odense University Hospital
STANLEY LEMESHOW, Ohio State University
J. SCOTT LONG, Indiana University
ROGER NEWSON, Imperial College, London
AUSTIN NICHOLS, Urban Institute, Washington DC
MARCELLO PAGANO, Harvard School of Public Health
SOPHIA RABE-HESKETH, Univ. of California–Berkeley
J. PATRICK ROYSTON, MRC Clinical Trials Unit,
London
PHILIP RYAN, University of Adelaide
MARK E. SCHAFFER, Heriot-Watt Univ., Edinburgh
JEROEN WEESIE, Utrecht University
IAN WHITE, MRC Biostatistics Unit, Cambridge
NICHOLAS J. G. WINTER, University of Virginia
JEFFREY WOOLDRIDGE, Michigan State University

Stata Press Editorial Manager

LISA GILMORE

Stata Press Copy Editors

DAVID CULWELL, SHELBI SEINER, and DEIRDRE SKAGGS

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

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

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

<http://www.stata-journal.com>

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

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

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

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

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

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

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

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

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

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



Copyright © 2015 by StataCorp LP

Copyright Statement: The *Stata Journal* and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

The articles appearing in the *Stata Journal* may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the *Stata Journal*.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the *Stata Journal*, in whole or in part, on publicly accessible websites, file servers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

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.

Regression analysis of censored data using pseudo-observations: An update

Morten Overgaard
Aarhus University
Aarhus, Denmark
moov@biostat.au.dk

Per K. Andersen
University of Copenhagen
Copenhagen, Denmark
pka@biostat.ku.dk

Erik T. Parner
Aarhus University
Aarhus, Denmark
parner@biostat.au.dk

Abstract. We present updated versions of the `stpsurv`, `stpci`, and `stpmean` commands, which were introduced in Parner and Andersen (2010, *Stata Journal* 10: 408–422), along with a new command, `stplost`. The commands generate pseudo-observations of the survival function, the cumulative incidence function under competing risks, the restricted mean survival-time function, and the cause-specific lost-lifetime function. The pseudo-observations can be used to assess the effects of covariates on their respective functions at different times by fitting generalized linear models to the pseudo-observations. The updated commands feature new options, an increase in computational speed, and the ability to handle survival data with delayed entry.

Keywords: st0202_1, `stpsurv`, `stpci`, `stpmean`, `stplost`, pseudovalues, time to event, survival analysis

1 Introduction

The concept of pseudo-observations in Andersen, Klein, and Rosthøj (2003) is based on the idea of the jackknife “leave-one-out” estimator and involves transforming a set of survival-time observations, possibly with right-censorings and delayed entry, into a set of pseudo-observations that can be used for regression analysis directly without further accounting for the possible right-censorings and delayed entries. Parner and Andersen (2010) introduced the commands `stpsurv`, `stpmean`, and `stpci` for generating these pseudo-observations for three different scenarios with the data subject to right-censoring. In this article, we present an update to these three commands, and we introduce a new one. This update was motivated by the desire to use the pseudo-observation approach on survival data with delayed entry. With the additional `stplost` command, pseudo-observations for the cause-specific lost-lifetime function can now be obtained as suggested in Andersen (2013). Also, the update includes a major revision of the Mata code that improves the computational speed as well as some new options that increase the flexibility of the commands.

2 The `stpsurv`, `stpmean`, `stpci`, and `stplost` commands

2.1 Syntax

```
stpsurv [if] [in], at(numlist) [generate(name) atnumbers failure
      after(time) replace]
```

```
stpmean [if] [in], at(numlist) [generate(name) atnumbers after(time)
      replace lostlifetime]
```

```
stpci [varname] [if] [in], at(numlist) [generate(name) atnumbers
      after(time) competingvalues(numlist) replace]
```

```
stplost [varname] [if] [in], at(numlist) [generate(name) atnumbers
      after(time) competingvalues(numlist) replace]
```

2.2 Description

The commands are used to generate the pseudo-observations in different situations: `stpsurv` is for the survival function, `stpmean` is for the restricted mean, `stpci` is for the cumulative incidence function under competing risks, and `stplost` is for the cause-specific lost-lifetime function. The time point or points at which the pseudo-observations are to be computed are specified using the `at()` option, which must be specified.

Because the commands are for use with `st` data, you must run `stset` before calling any of the commands.

Weights cannot be specified in the commands, but frequency weights used in the preliminary `stset` command will be preserved and used when generating the pseudo-observations. Other weight types that may have been used in the `stset` command are ignored.

With `stpci` for generating pseudo-observations of the cumulative incidence function under competing risks, values of the failure variable other than the ones specified in the `stset` command (the events) and 0 (taken as the censorings) will be regarded as competing risks by default. If competing-risks indicators are kept in another variable, this can be specified using the optional *varname*.

The pseudo-observations are by default stored in the variable `pseudo` when one time point is specified and in the variables `pseudo1`, `pseudo2`, ..., `pseudok` when *k* time points are specified. The variable name can be changed using the `generate(name)` option.

2.3 Options

`at(numlist)` determines the time points at which to compute the pseudo-observations. Negative values are ignored. `at()` is required.

`generate(name)` specifies a variable name for the pseudo-observations. The default is `generate(pseudo)`. If two or more time points are specified, the variable names will by default be *name1*, *name2*, and so on.

`atnumbers` changes the numbering of the created variables if there are two or more time points specified. The numbers will match the time points specified in `at(numlist)` instead of the default $1, 2, \dots, k$. To ensure that the result is a valid variable name, a decimal like 1.9 will be represented as 1_9.

`failure` is used with `stpsurv` and creates pseudo-observations for the (partial) failure-probability function instead of the survival function.

`after(time)` excludes observations that exit before time point *time* and thus considers mortality only after that time point in the calculation. It is equivalent to specifying `if >= time` in the optional `if` criteria.

`competingvalues(numlist)` is used with `stpci` and `stplost` and determines the values of *varname* that are regarded as competing events. If the optional *varname* has not been specified, the variable used as the failure variable in the `stset` command will be used. When *varname* has been specified, the default is `competingvalues(1)`, and when *varname* is not specified, the default is to take all values of the failure variable except 0 (considered censorings) and the ones marking events as set in the `stset` command.

`replace` specifies that existing variables be replaced with the newly created variables without error.

`lostlifetime` is used with `stpmean` and creates pseudo-observations for the lost-lifetime function instead of the restricted mean function.

3 Changes and comments on technical aspects

The commands will work with Stata 11 and later. Almost any use of the earlier versions of the commands will be compatible with the newer versions.

Changes from previous versions of the commands include the following:

1. The `at()` option is now more liberal and can handle a nonascending *numlist*. Also, time points outside the range of the survival data can now be specified—if this is done, you will receive a note instead of an error.
2. The competing variable and values for `stpci` can now be specified more freely using the optional *varname* and the `competingvalues()` option.
3. The options `atnumbers` and `replace` have been added to ease the variable handling.

4. The `after()` option has been added to make it easier to focus analyses on mortality after a certain time point.
5. Weight types other than frequency weights are ignored. The commands are designed to handle only frequency weights.
6. A `conditional` option for `stpmean` has been removed. The option gave a conditional mean survival time as an alternative to the restricted mean survival time. Because the conditional mean survival time was conditional on having an event before the specified time point, the option was in retrospect considered to violate the survival analysis principles of sticking to this world and not conditioning on the future (see Andersen and Keiding [2012]).

Computational efficiency increased considerably by making the Mata code more vectorized and limiting the number of calculations performed. That the Kaplan–Meier curve jumps only in time points with one or more events implies that the effect of an individual on the Kaplan–Meier curve is determined by the first event time after an individual’s entry and the last event time before the individual has left the study. The latter will be the individual’s own event time if he or she has one. Individuals for whom these two time points are the same will affect the Kaplan–Meier curve similarly and will always have the same Kaplan–Meier-based pseudo-observation. Using this observation means that the computational time depends much more on the number of distinct event times than on the number of individuals in the study. Thus the increase in computational speed from the previous version will grow with the rarity of the event under consideration and thereby make it feasible to use the pseudo-observation method on larger datasets with a relatively rare event. An indication of the reduction in execution time is given by table 1, which shows the observed reduction for `stpsurv` in various simulated examples. Similar results can be expected for `stpmean` and `stpci`.

Table 1. Reduction in execution time from the previous version of `stpsurv`

<i>n</i>	<i>p</i>			
	0.1	0.2	0.4	0.8
500	95.1%	90.9%	87.7%	78.3%
1000	96.9%	96.0%	92.2%	87.1%
2000	98.4%	97.9%	95.8%	91.5%
4000	99.4%	98.8%	97.2%	92.3%

Note: Simulation of n observations with probability p of having an event at the end of the time under observation. The time under observation was taken from a uniform distribution on $[0, 1]$. The `stpsurv` command was asked to calculate pseudo-observations at time 1. Each configuration was run 10 times, and the reduction in average execution time is reported.

4 Pseudo-observations

In this section, we briefly describe what pseudo-observations are and how they are calculated. More details can be found in, for example, Andersen and Pohar Perme (2010).

To calculate a set of pseudo-observations, we need to settle on some parameter of interest. Here we deal with the values

- $S(t)$, the survival function at some time point, t ;
- $\mu(t)$, the restricted mean survival time, restricted to time t ;
- $\text{CIP}_i(t)$, the cumulative incidence probability (CIP) function for cause i at time t ; and
- $L_i(0, t)$, the cause-specific (i) lost-lifetime function at time t .

If we let θ denote one of the above, we can use the Kaplan–Meier or Aalen–Johansen method to obtain an estimate, $\hat{\theta}$, of θ . If we let $\hat{\theta}_{-i}$ denote the similar estimate obtained by using all the data except for the i th observation, the i th pseudo-observation is

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}$$

where n is the total number of observations.

The actual estimators used can be described by the formulas

$$\hat{S}(t) = \prod_{t_j \leq t} \frac{Y_j - d_j}{Y_j}$$

where t_1, \dots, t_D are the distinct event times, Y_j is the number we know to be at risk at time t_j , and d_j is the number of observed events at time t_j , as well as

$$\begin{aligned} \hat{\mu}(t) &= \int_0^t \hat{S}(u) du \\ \widehat{\text{CIP}}_i(t) &= \sum_{t_j \leq t} \frac{d_{ij}}{Y_j} \hat{S}(t_j-) \end{aligned}$$

where d_{ij} is the number of events of cause i at time t_j and $\hat{S}(t_j-)$ is the left-hand limit of \hat{S} at t_j ,

$$\hat{L}_i(0, t) = \int_0^t \widehat{\text{CIP}}_i(u) du$$

As long as the Y_j s are the numbers observed to be at risk at their respective time points, t_j , these estimators are reasonable estimators even in a situation with delayed entry.

Once the pseudo-observations have been calculated, they can be used in a generalized linear model for assessing the effects of covariates. Graw, Gerds, and Schumacher (2009)

proved that this approach will yield asymptotically unbiased and normal distributed estimates under the assumption that censoring is independent of covariates.

This approach has been applied to real and simulated data in a series of articles. Klein et al. (2007) studied the pseudo-observation approach for the survival function. Andersen, Hansen, and Klein (2004) found promising results for the pseudo-observation approach for the restricted mean survival time. Klein and Andersen (2005) studied the pseudo-observation approach for the cumulative incidence function under competing risks using Monte Carlo simulation methods and found it to be a viable method. Recently, Hansen, Andersen, and Parner (2014) studied the necessary number of events per explanatory variable for the pseudo-observation approach on the survival function in a simulation study that recommended using more than 10 events per variable for the study of risk difference and more than 15 events per variable for the study of relative risk. This study also found a better coverage probability for the pseudo-observation method in the study of risk difference on a small sample compared with the coverage probability obtained by comparing Kaplan–Meier estimates for the two groups.

5 Examples

5.1 Diabetics in the county of Funen

In this example, we use data from an epidemiological study of diabetes mellitus in Denmark (Green et al. 1981). A total of 1,499 diabetics alive on 1 July 1973 and living in the county of Funen were identified and followed up on their vital status until 1 January 1982. We have changed the data slightly by adding or subtracting a few days at random from both the birth date and the death date, and any deaths occurring after 1 January 1982 are considered censorings on that date.

Because age is such an important risk factor for death, we may want to use age as the time variable in our analyses, making the entry in the study on 1 July 1973 a “delayed entry”.

Suppose we wanted to compare survival between men and women in this group of diabetics in Funen. Let’s set up the survival data (`stset` the data) as described above.

```

. use diabetes_rev.dta
. stset date, failure(status) origin(birthdate) enter(time mdy(7,1,1973))
> id(id) scale(365.25)

      id:  id
failure event:  status != 0 & status < .
obs. time interval:  (date[_n-1], date]
enter on or after:  time mdy(7,1,1973)
exit on or before:  failure
      t for analysis:  (time-origin)/365.25
      origin:  time birthdate

```

```

1499  total observations
      0  exclusions

```

```

1499  observations remaining, representing
1499  subjects
478   failures in single-failure-per-subject data
10601.692  total analysis time at risk and under observation
              at risk from t =          0
              earliest observed entry t = 2.844627
              last observed exit t = 96.22177

```

Here `date` is the date of death or censoring at the latest at the study end, 1 January 1982. To avoid the initial period with only a few individuals at risk, we will limit ourselves to individuals who survived to adulthood (to 18 years). Also, let's say that we are primarily interested in survival at age 75. The `sts list` command can give us the Kaplan–Meier failure estimates at age 75 for each sex for those who had not exited at age 18.

```

. sts list if _t >= 18, failure by(sex) at(0 75)

      failure _d:  status
analysis time _t:  (date-origin)/365.25
      origin:  time birthdate
enter on or after:  time mdy(7,1,1973)
      id:  id

```

	Time	Beg. Total	Fail	Failure Function	Std. Error	[95% Conf. Int.]	
female							
	0	0	0	1.0000	.	.	.
	75	79	135	0.8012	0.0280	0.7438	0.8528
male							
	0	0	0	1.0000	.	.	.
	75	48	188	0.9097	0.0162	0.8747	0.9382

Note: Failure function is calculated over full data and evaluated at indicated times; it is not calculated from aggregates shown at left.

The estimated risk difference will be $0.9097 - 0.8012 = 0.1085$ with a 95% confidence interval of $[0.0451, 0.1719]$ based on the standard error $\sqrt{0.0280^2 + 0.0162^2} = 0.0323$. Similarly, the estimated relative risk will be $0.9097/0.8012 = 1.135$, that is, a 13.5% increased risk for the male sex. A confidence interval is not as easily available, but one based on the delta method can be calculated.

In the pseudo-observation approach, we start by calculating the pseudo-observations for the failure function at age 75. We want to consider mortality after 18 years only and specify the `after(18)` option.

```
. stpsurv, at(75) failure after(18)
Pseudo-observations for the failure probability function.
Computing pseudo-observations (progress dots indicate percent completed).
----- 1 ----- 2 ----- 3 ----- 4 ----- 5
..... 50
..... 100
Generated variable: pseudo.
```

We can now estimate the risk difference at age 75 by running the `glm` command on the pseudo-observations with sex as an explanatory variable.

```
. glm pseudo i.sex, vce(robust)
Iteration 0:   log pseudolikelihood = -1299.7365
Generalized linear models               No. of obs   =       1,470
Optimization      : ML                  Residual df   =       1,468
                                      Scale parameter =       .343638
Deviance          = 504.4605253          (1/df) Deviance =       .343638
Pearson           = 504.4605253          (1/df) Pearson  =       .343638
Variance function: V(u) = 1              [Gaussian]
Link function     : g(u) = u              [Identity]
                                      AIC              =       1.77107
Log pseudolikelihood = -1299.736499      BIC              =      -10201.69
```

pseudo	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
sex						
male	.1057376	.0305828	3.46	0.001	.0457964	.1656789
_cons	.8091827	.0219744	36.82	0.000	.7661138	.8522517

The variance of a pseudo-observation-based estimate is usually estimated using a sandwich estimator, which is why the option `vce(robust)` has been specified. We see the risk difference is estimated by 0.1057 with a confidence interval of [0.0458, 0.1657]—not far from the results from the `sts list` command.

If we do the same thing with the log-link function (and the `eform` option specified), we will obtain an estimate of the relative risk.

```
. glm pseudo i.sex, link(log) eform vce(robust)
(output omitted)
```

```
Generalized linear models          No. of obs      =       1,470
Optimization      : ML              Residual df    =       1,468
                                   Scale parameter =    .343638
Deviance          = 504.4605253      (1/df) Deviance =    .343638
Pearson           = 504.4605253      (1/df) Pearson  =    .343638
Variance function: V(u) = 1          [Gaussian]
Link function     : g(u) = ln(u)     [Log]
                                   AIC          =    1.77107
                                   BIC          =   -10201.69
Log pseudolikelihood = -1299.736499
```

pseudo	exp(b)	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
sex						
male	1.130672	.0404199	3.44	0.001	1.054162	1.212735
_cons	.8091827	.0219744	-7.80	0.000	.7672399	.8534185

The estimate of the relative risk is 1.131, and this time we get the confidence interval [1.054, 1.213]—that is, we get between 5.4% and 21.3% increased risk of dying before age 75 given survival to age 18 for the male diabetics in Funen.

To adjust for a potential difference due to the time of diagnosis, we can include, for example, a variable on the decade of diagnosis.

```
. glm pseudo i.sex i.diagnosis_decade, link(log) eform vce(robust)
(output omitted)
```

```
Generalized linear models          No. of obs      =       1,470
Optimization      : ML              Residual df    =       1,463
                                   Scale parameter =    .3434418
Deviance          = 502.4553284      (1/df) Deviance =    .3434418
Pearson           = 502.4553284      (1/df) Pearson  =    .3434418
Variance function: V(u) = 1          [Gaussian]
Link function     : g(u) = ln(u)     [Log]
                                   AIC          =    1.77389
                                   BIC          =   -10167.23
Log pseudolikelihood = -1296.809101
```

pseudo	exp(b)	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
sex						
male	1.138027	.041397	3.55	0.000	1.059715	1.222126
diagnosis_d~e						
1930	1.00076	.1789723	0.00	0.997	.7048619	1.420875
1940	1.01044	.1670864	0.06	0.950	.7307276	1.397221
1950	.9513142	.1512878	-0.31	0.754	.6965577	1.299244
1960	.9050087	.142623	-0.63	0.527	.664522	1.232526
1970	.8912134	.1436501	-0.71	0.475	.6498019	1.222313
_cons	.8622548	.1357217	-0.94	0.346	.6333646	1.173863

In this example, including the additional variable did not make a huge difference. However, we see that by using the pseudo-observations, we can now easily make regression analyses of, for example, the failure function at one or more time points of our choosing.

5.2 Follicular cell lymphoma study

Here we use data on records of 541 patients diagnosed with an early-stage follicular-type lymphoma registered for treatment at the Princess Margaret Hospital in Toronto. These data were described in Pintilie (2006) and made available on her webpage. Important variables include **age** (in years), **hgb** (hemoglobin, g/L), **clinstg** (clinical stage, I or II), **ch** (treated with chemotherapy), and **dftime** (time from diagnosis to first failure: no response, relapse, or death). An event indicator, **event**, was generated using the variables **resp** (response after treatment: CR for complete response, NR for no response); **relnsite** (site of relapse: blank if no relapse); and **dfcens** (indicates an event: 1 for failure, 0 for censoring) with encoding (0 for censoring, 1 for relapse-free death, 2 for relapse, and 3 for no response to treatment).

```
. generate event = 1*(dfcens==1 & relsite==" " & resp=="CR")
> + 2*(dfcens==1 & relsite!=" " & resp=="CR") + 3*(resp=="NR")
```

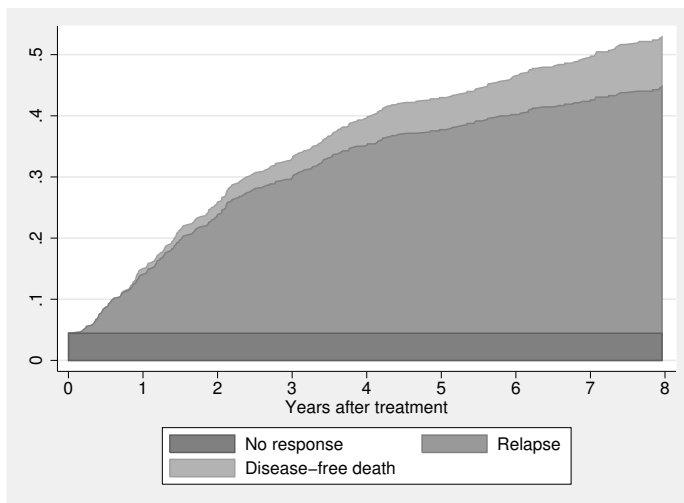


Figure 1. Stacked cumulative incidence proportions for the three event types of the follicular cell lymphoma study for eight years of follow-up

The event of no response to treatment is determined immediately and therefore does not compete with the other events over time. In figure 1, we can see that relapse is the dominating event in the first eight years after treatment. If we are interested in the first year after treatment, the event of disease-free death can be ignored without much error, but let's say we are interested in what happens up to five years of follow-up. In

this case, it would be appropriate to treat disease-free death and relapse as competing risks. We treat the three event types as competing risks.

We are mainly interested in the effect of chemotherapy on the occurrence of relapse, and we want an adjusted estimate of that effect, adjusting for certain variables that could be important. We generate the following variables for use with `glm`:

```
. generate chemo = (ch=="Y")
. generate age50a5 = (age-50)/5
. generate hgb140a10 = (hgb-140)/10
```

We must first use the `stset` command to make relapse the event of interest.

```
. stset dftime, failure(event==2) id(stnum)
      id:  stnum
      failure event:  event == 2
obs. time interval:  (dftime[_n-1], dftime]
exit on or before:  failure
```

```
541  total observations
    0  exclusions
```

```
541  observations remaining, representing
541  subjects
248  failures in single-failure-per-subject data
4000.041  total analysis time at risk and under observation
                                             at risk from t =      0
                                             earliest observed entry t =    0
                                             last observed exit t = 31.10198
```

We calculate the pseudo-observations for the cumulative incidence proportion at five years and use the pseudo-observations in a `glm` with our variables on chemotherapy (indicator), age, hemoglobin level, and clinical stage to get an estimate of the parameter associated with chemotherapy adjusted for the other variables. We model the effect of the variables using a log-link function; hence, our estimates are relative to the base level.

```
. stpci, at(5) generate(pseudo_ci)
Pseudo-observations for the cumulative incidence function.
Competing risks: event = 1 3.
Computing pseudo-observations (progress dots indicate percent completed).
-----| 1 -----| 2 -----| 3 -----| 4 -----| 5
..... 50
..... 100
Generated variable: pseudo_ci.
. glm pseudo_ci i.chemo c.age50a5 c.hgb140a10 i.clinstg, vce(robust)
> link(log) eform noheader nolog
```

pseudo_ci	exp(b)	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.chemo	.4794364	.0995162	-3.54	0.000	.319189	.7201353
age50a5	1.044187	.0236496	1.91	0.056	.998848	1.091583
hgb140a10	1.065364	.0450557	1.50	0.134	.9806172	1.157435
2.clinstg	1.382896	.1844996	2.43	0.015	1.064698	1.796193
_cons	.3145069	.0320075	-11.37	0.000	.257634	.3839346

As we can see, the `stpci` command automatically uses events 1 and 3 as competing risks. The parameter estimate of 0.479 (0.319–0.720) means the ones who are treated with chemotherapy have a probability of relapse within the first 5 years after treatment that is reduced by 52.1% (28.0%–62.1%) compared with the ones who were not treated with chemotherapy when adjusting for the other variables.

Sometimes, the cumulative incidence proportion is considered inappropriate or too abstract as a measure of the rate of occurrence of an event in a competing-risks scenario. Another measure, as suggested by Andersen (2013), that can be considered more concrete is the expected (disease-free) lifetime lost because of a given event up to a given time point. We calculate pseudo-observations for the lost-lifetime function at five years and use these in a `glm` similar to the one we used for the cumulative incidence proportion, this time using the identity-link function.

```
. stplot, at(5) generate(pseudo_lost)
Pseudo-observations for the lost life time function.
Competing risks: event = 1 3.
Computing pseudo-observations (progress dots indicate percent completed).
-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 1 | 2 | 3 | 4 | 5 |
..... 50
..... 100
Generated variable: pseudo_lost.
. glm pseudo_lost i.chemo c.age50a5 c.hgb140a10 i.clinstg, vce(robust)
> link(id) noheader nolog
```

pseudo_lost	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]	
1.chemo	-.7582157	.1399544	-5.42	0.000	-1.032521	-.4839103
age50a5	.0320911	.0246744	1.30	0.193	-.0162698	.080452
hgb140a10	.0381067	.0439318	0.87	0.386	-.0479981	.1242114
2.clinstg	.4723689	.157472	3.00	0.003	.1637294	.7810085
_cons	.9945779	.0992169	10.02	0.000	.8001163	1.189039

The estimate of -0.758 (with confidence interval $[-1.033, -0.484]$) means the ones treated with chemotherapy can expect to lose about three quarters of a year less disease-free lifetime in the first 5 years after treatment than the ones not treated with chemotherapy when adjusting for the other variables.

6 Concluding remarks

We have presented updated versions of the `stpsurv`, `stpci`, and `stpmean` commands along with the new command `stplot`. These updates make computing pseudo-observations a straightforward task for Stata users. We believe that the pseudo-observation method is a versatile and proper tool for regression analysis of censored time-to-event data and that practitioners will benefit from embracing the pseudo-observation method.

7 References

- Andersen, P. K. 2013. Decomposition of number of life years lost according to causes of death. *Statistics in Medicine* 32: 5278–5285.
- Andersen, P. K., M. G. Hansen, and J. P. Klein. 2004. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis* 10: 335–350.
- Andersen, P. K., and N. Keiding. 2012. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine* 31: 1074–1088.
- Andersen, P. K., J. P. Klein, and S. Rosthøj. 2003. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* 90: 15–27.
- Andersen, P. K., and M. Pohar Perme. 2010. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research* 19: 71–99.
- Graw, F., T. A. Gerds, and M. Schumacher. 2009. On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* 15: 241–255.
- Green, A., M. Hauge, N. V. Holm, and L. L. Rasch. 1981. Epidemiological studies of diabetes mellitus in Denmark. II. A prevalence study based on insulin prescriptions. *Diabetologia* 20: 468–470.
- Hansen, S. N., P. K. Andersen, and E. T. Parner. 2014. Events per variable for risk differences and relative risks using pseudo-observations. *Lifetime Data Analysis* 20: 584–598.
- Klein, J. P., and P. K. Andersen. 2005. Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* 61: 223–229.
- Klein, J. P., B. Logan, M. Harhoff, and P. K. Andersen. 2007. Analyzing survival curves at a fixed point in time. *Statistics in Medicine* 26: 4505–4519.
- Parner, E. T., and P. K. Andersen. 2010. Regression analysis of censored data using pseudo-observations. *Stata Journal* 10: 408–422.
- Pintilie, M. 2006. *Competing Risks: A Practical Perspective*. Chichester, UK: Wiley.

About the authors

Morten Overgaard has an MSc in statistics and is a research assistant at Aarhus University, where he is supervised by coauthor Erik T. Parner.

Per K. Andersen has a PhD in statistics and a DMSc degree in biostatistics, both from the University of Copenhagen. He is a professor of biostatistics at the University of Copenhagen. His main research fields are time-to-event analysis and statistical methods in epidemiology.

Erik T. Parner has a PhD in statistics from Aarhus University. He is a professor of biostatistics at Aarhus University. His research fields are time-to-event analysis, statistical methods in epidemiology and genetics, and the etiology and changing prevalence of autism.