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 BRICK, New York University
RINO BELLOCCO, Karolinska Institutet, Sweden, and
University of Milano-Bicocca, Italy
MAARTEN L. BUSS, WZB, 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 Lauritzen, 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 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, 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
**Subscription rates** listed below include both a printed and an electronic copy unless otherwise mentioned.

<table>
<thead>
<tr>
<th></th>
<th>U.S. and Canada</th>
<th>Elsewhere</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Printed &amp; electronic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-year subscription</td>
<td>$ 98</td>
<td>$138</td>
</tr>
<tr>
<td>2-year subscription</td>
<td>$165</td>
<td>$245</td>
</tr>
<tr>
<td>3-year subscription</td>
<td>$225</td>
<td>$345</td>
</tr>
<tr>
<td>1-year student subscription</td>
<td>$ 75</td>
<td>$ 99</td>
</tr>
<tr>
<td>1-year institutional subscription</td>
<td>$245</td>
<td>$285</td>
</tr>
<tr>
<td>2-year institutional subscription</td>
<td>$445</td>
<td>$525</td>
</tr>
<tr>
<td>3-year institutional subscription</td>
<td>$645</td>
<td>$765</td>
</tr>
<tr>
<td><strong>Electronic only</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-year subscription</td>
<td>$ 75</td>
<td>$ 75</td>
</tr>
<tr>
<td>2-year subscription</td>
<td>$125</td>
<td>$125</td>
</tr>
<tr>
<td>3-year subscription</td>
<td>$165</td>
<td>$165</td>
</tr>
<tr>
<td>1-year student subscription</td>
<td>$ 45</td>
<td>$ 45</td>
</tr>
</tbody>
</table>

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


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


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 © 2013 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.
Flexible parametric illness-death models

Sally R. Hinchliffe  
Department of Health Sciences  
University of Leicester  
Leicester, UK  
srh20@leicester.ac.uk

David A. Scott  
Oxford Outcomes Ltd  
Oxford, UK  
david.scott@oxfordoutcomes.com

Paul C. Lambert  
Department of Health Sciences  
University of Leicester  
Leicester, UK  
and  
Department of Medical Epidemiology and Biostatistics  
Karolinska Institutet  
Stockholm, Sweden  
paul.lambert@le.ac.uk

Abstract. It is usual in time-to-event data to have more than one event of interest, for example, time to death from different causes. Competing risks models can be applied in these situations where events are considered mutually exclusive absorbing states. That is, we have some initial state—for example, alive with a diagnosis of cancer—and we are interested in several different endpoints, all of which are final. However, the progression of disease will usually consist of one or more intermediary events that may alter the progression to an endpoint. These events are neither initial states nor absorbing states. Here we consider one of the simplest multistate models, the illness-death model. stpm2illd is a postestimation command used after fitting a flexible parametric survival model with stpm2 to estimate the probability of being in each of four states as a function of time. There is also the option to generate confidence intervals and transition hazard functions. The new command is illustrated through a simple example.

Keywords: st0316, illdprep, stpm2illd, survival analysis, multistate models, flexible parametric models

1 Introduction

It is usual in time-to-event data to have more than one event of interest, for example, time to death from different causes. If we treat these events as mutually exclusive endpoints where the occurrence of an event is final, then we can apply a competing risks model (Prentice et al. 1978; Colzani et al. 2011; Hinchliffe and Lambert 2013a,b). These endpoints are known as absorbing states, and we model the time to each of these from some initial state, for example, alive with a diagnosis of cancer. However, the progression of disease will usually consist of one or more intermediary events that may
alter the progression to an endpoint (Putter, Fiocco, and Geskus 2007). These events
cannot be classified as initial states or absorbing states and so are known as transient
states or intermediate states.

Illness-death models are a special case of multistate models, where individuals start
out healthy and then may become ill and go on to die. In theory, some patients may
recover from an illness and become healthy again (Andersen, Abildstrom, and Rosthoj
2002). This is known as a bidirectional illness-death model. We will consider only the
unidirectional model as illustrated in figure 1.

The two main measures of interest for analyses of this type are the transition hazards
and the probability of being in each state as a function of time. The transition hazards
can inform us about the impact of risk factors on rates of illness and disease or mortality.
Additionally, the probabilities of being in each state provide an absolute measure on
which to base prognosis and clinical decisions (Koller et al. 2012). The purpose of this
article is to explain how to set up the data using illdprep in a format that allows
flexible parametric survival models (stpm2) to estimate transition hazards. Using the
postestimation command stpm2illd, we can then obtain both the probability of being
in each state as a function of time and the confidence intervals for each.

Figure 1. Unidirectional illness-death model
2 Methods

Figure 1 shows a graphical representation of a unidirectional illness-death model. The states are represented with a box and given a number from one to four. The transitions are represented by arrows going from one state to another. In total, there are three transitions labeled from one to three. We represent a transition from state $i$ to $j$ by $i \rightarrow j$; therefore, the transition hazards are denoted on the diagram as $\alpha_{13}$, $\alpha_{12}$, and $\alpha_{24}$ (Putter, Fiocco, and Geskus 2007). If $T$ denotes the time of reaching state $j$ from state $i$, we denote the hazard rate (transition intensity) of the $i \rightarrow j$ transition by

$$
\alpha_{ij} = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}
$$

Currently, most applications of illness-death models involve the Cox model. However, we are interested in parametric estimates and so advocate the use of the flexible parametric survival model, first proposed by Royston and Parmar (2002). The approach uses restricted cubic spline functions to model the baseline log cumulative hazard. It has the advantage over other well-known models such as the Cox model because it produces smooth predictions and can be extended to incorporate complex time-dependent effects, again through the use of restricted cubic splines. The Stata implementation of the model using stpm2 is described in detail elsewhere (Lambert and Royston 2009).

The transition hazard rates in (1) can be obtained from the flexible parametric survival model. This could be done by fitting separate models for each of the three transitions, but this would not allow for shared parameters. It is possible to fit one model for all three transitions simultaneously by stacking the data so that each individual patient has up to three rows of data, dependent on how many transitions each patient is at risk of.

Table 1 shows four cancer patients of varying ages who are all at risk of both relapse of their cancer and death. Relapse can be considered an intermediary event, whereas death is final and thus an absorbing state. Patient 1, aged 44, is at risk of both relapse and death for 2.4 years until the patient relapses and goes on to die after 7.6 years. Patient 2, aged 68, is at risk of both relapse and death for 9 years until the patient dies and is no longer at risk of relapse. Patient 3, aged 52, is at risk of both relapse and death until the patient is censored at 6.1 years. Finally, patient 4, aged 38, is at risk of both relapse and death for 4.6 years until the patient relapses and is at risk of death until being censored at 13.8 years.

To model all three transitions simultaneously, we need to set up the data as shown in table 2. The data have been expanded so that each patient now has up to three rows of data. As shown in figure 1, transition 1 goes from alive and well to dead, transition 2 goes from alive and well to ill, and transition 3 goes from ill to dead. Patient 1 is at risk of both relapse (state 2) and death (state 3) for 2.4 years when the patient relapses. The patient is then at risk of death with relapse (state 4) from 2.4 years to 7.6 years, when he or she dies. Patient 2 is at risk of both relapse (state 2) and death (state 3) for 9 years until the patient dies and is no longer at risk of relapse. Because patient 2 never experienced a relapse, the patient is never at risk of experiencing state 4. Therefore, in
Flexible parametric illness-death models

the expanded data, he or she has only two rows of data. Patient 3 is at risk of both relapse (state 2) and death (state 3) for 6.1 years when the patient is censored from the study. Again, because patient 3 never experienced a relapse, the patient is never at risk of experiencing transition 3 and thus has only two rows of data. Finally, patient 4 is at risk of both relapse (state 2) and death (state 3) for 4.6 years when he or she relapses. The patient is then at risk of death with relapse (state 4) from 4.6 years to 13.8 years when the patient is censored.

Table 1. Standard dataset with relapse and survival times (years) for four patients

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Relapse time</th>
<th>Relapse indicator</th>
<th>Survival time</th>
<th>Death indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44</td>
<td>2.4</td>
<td>1</td>
<td>7.6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>9.0</td>
<td>0</td>
<td>9.0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>6.1</td>
<td>0</td>
<td>6.1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>4.6</td>
<td>1</td>
<td>13.8</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2. Expanded dataset with transition indicators and start and stop times (years) for four patients

<table>
<thead>
<tr>
<th>ID</th>
<th>Age</th>
<th>Trans 1</th>
<th>Trans 2</th>
<th>Trans 3</th>
<th>Status</th>
<th>Start</th>
<th>Stop</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.4</td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2.4</td>
</tr>
<tr>
<td>1</td>
<td>44</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2.4</td>
<td>7.6</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9.0</td>
</tr>
<tr>
<td>2</td>
<td>68</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9.0</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.1</td>
</tr>
<tr>
<td>3</td>
<td>52</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6.1</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.6</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4.6</td>
</tr>
<tr>
<td>4</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4.6</td>
<td>13.8</td>
</tr>
</tbody>
</table>

The transition hazard rates can be transformed into the probability of being in each of the four states (state occupation probabilities) through the following relationships. Notice that as in the competing risks setting, there is not a one-to-one correspondence between the transition hazards and the transition probabilities: the latter is a function of multiple transition hazards.
The probability of being alive and well will depend on both the transition rate from alive to dead \([\alpha_{13}(t)]\) and the transition rate from alive to relapse \([\alpha_{12}(t)]\). An individual needs to have survived both death (state 3) and illness (state 2) to remain in the state representing alive and well. This is essentially the survival probability where both death and illness are considered events.

\[
P(\text{alive and well at time } t) = \exp\left\{-\int_{0}^{t} \alpha_{13}(s) + \alpha_{12}(s) ds\right\} \tag{2}
\]

When estimating the probability of being alive with illness, we have to consider not only the probability of getting ill but also the probability of remaining alive with the illness (that is, of not moving to state 4). The probability of being ill is a function of the transition hazard from alive (state 1) to ill (state 2) and the probability of being alive and well from (2). The probability of remaining alive with the illness (that is, staying in state 2) is the survival function for the transition from ill to death (transition 3 in figure 1).

\[
P(\text{alive with illness at time } t) = \int_{0}^{t} (\text{ill at time } s) \times P(\text{survive with illness from } s \text{ to } t) ds = \alpha_{12}(s) \exp\left\{-\int_{0}^{s} \alpha_{13}(u) + \alpha_{12}(u) du\right\} \times \exp\left\{-\int_{s}^{t} \alpha_{24}(u) du\right\} ds \tag{3}
\]

The probability of dying without illness is a function of the transition hazard from alive (state 1) to dead (state 3) and the probability of being alive and well from (2).

\[
P(\text{dead without illness at time } t) = \int_{0}^{t} \alpha_{13}(s) \exp\left\{-\int_{0}^{s} \alpha_{13}(u) + \alpha_{12}(u) du\right\} ds \tag{4}
\]

Finally, the probability of dying with illness can be estimated by subtracting the probability of being in each of the other three states from 1.

\[
P(\text{dead with illness at time } t) = 1 - P(\text{alive and well at time } t) - P(\text{ill at time } t) - P(\text{dead without illness at time } t) \tag{5}
\]

To get the overall probability of death at time \(t\), we add \(P(\text{dead without illness at time } t)\) and \(P(\text{dead with illness at time } t)\). Confidence intervals can be calculated for each of these probabilities using the delta method (Carstensen 2006; Lambert et al. 2010).
3 The illdprep command

The illdprep command is used before stset and stpm2 to set the data up in the format needed for illness-death models as shown in table 2 in section 2.

3.1 Syntax

ildprep, id(varlist) statevar(varlist) statetime(varlist) [status(varname) 
transname(varlist) addtime(real)]

3.2 Options

id(varlist) specifies the name of the ID variable in the dataset. Before the command is used, each ID number should have just one row of data. The command will expand the data so that each ID number will have up to three rows of data. id() is required.

statevar(varlist) specifies the names of the two event-indicator variables needed to split the data. As demonstrated in figure 1 and table 2 an indicator variable will be needed to specify whether a patient has become ill and whether a patient has died. Because death is a final absorbing state, this must come last in the varlist. So, for example, if we were interested in relapse and death and our event-indicator variables were relapse and dead, then we would specify statevar(relapse dead) in that order. statevar() is required.

statetime(varlist) specifies the names of the two event-time variables. The variables should be input in the order that corresponds to statevar(varlist). So if our event-time variables were relapsetime and survtime, then we would specify statetime(relapsetime survtime) in that order to correspond with the example given for statevar(varlist). statetime() is required.

status(varname) allows the user to specify the name of the newly generated status variable as shown in table 2.

transname(varlist) allows the user to specify the names of the newly generated transition indicators. The default for these is trans1, trans2, and trans3. The user must specify these in the order that corresponds with figure 1 varlist must contain three variable names.

addtime(real) specifies an amount to add to the death time when event times are tied. For example, if a patient both relapses and dies at the same time in the data, then the user could add 0.1 to the death time so that the stset command does not drop the third transition. The specified value will obviously depend on the time units in the data.
4 The \texttt{stpm2illd} command

The \texttt{stpm2illd} command is a postestimation command used after \texttt{stpm2} to obtain the predictions given in (2), (3), (4), and (5) in section 2. The names specified in \texttt{newvarlist} coincide with the order of the transitions entered in the options.

4.1 Syntax

\begin{verbatim}
\texttt{stpm2illd newvarlist, trans1(varname # [varname # ...]) trans2(varname # [varname # ...]) trans3(varname # [varname # ...]) [obs(integer) ci mint(real) maxt(real) timename(varname) hazard hazname(varlist) combine]}
\end{verbatim}

4.2 Options

\texttt{trans1(varname # [varname # ...]) ... trans3(varname # [varname # ...])} requests that the covariates specified by the listed \texttt{varname} be set to \texttt{#} when predicting the hazards for each transition. The transition numbers correspond to those in the diagram above. Therefore, \texttt{trans1()} relates to the transition from alive to dead, \texttt{trans2()} relates to the transition from alive to ill, and \texttt{trans3()} relates to the transition from ill to dead. \texttt{trans1()}, \texttt{trans2()}, and \texttt{trans3()} are required.

\texttt{obs(integer)} specifies the number of observations (of time) to predict for. The default is \texttt{obs(1000)}. Observations are evenly spread between the minimum and maximum value of follow-up time. Note: Because the command uses numerical integration, if the number of specified observations is too small, then it may result in biased estimates.

\texttt{ci} calculates a 95\% confidence interval for the probabilities of being in each state and stores the confidence limits in \texttt{prob\_newvar\_lci} and \texttt{prob\_newvar\_uci}.

\texttt{mint(real)} specifies the minimum value of follow-up time. The default is set as the minimum event time from \texttt{stset}.

\texttt{maxt(real)} specifies the maximum value of follow-up time. The default is set as the maximum event time from \texttt{stset}.

\texttt{timename(varname)} is the name given to the time variable used for predictions. The default is \texttt{timename(newt)}. Note that this is the variable for time that needs to be used when plotting curves for the transition hazards and probabilities.

\texttt{hazard} predicts the hazard function for each transition.

\texttt{hazname(varlist)} allows the user to specify the names for the transition hazards if the \texttt{hazard} option is chosen. These will then be stored in variables called \texttt{h\_var}. The default is \texttt{hazname(trans1 trans2 trans3)}, which cause variables \texttt{h\_trans1}, \texttt{h\_trans2}, \texttt{h\_trans3} to be created. \texttt{varlist} must contain three variable names.
Flexible parametric illness-death models

combine allows the user to combine the probabilities of being in states 3 and 4 to give the overall probability of death. If this option is specified, then the user only needs to specify three names in newvarlist. The last name given in the list should correspond to the combined probability of states 3 and 4. So, for example, if we write alive ill dead in the newvarlist, then the probability of being in each state as a function of time will be stored as prob_alive, prob_ill, and prob_dead.

5 Example

The Rotterdam breast cancer data used in this example are taken from Royston and Lambert (2011). Download the data at http://www.stata-press.com/data/fpsaus.html. The data contain information on 2,982 patients with primary breast cancer. Both the time to relapse and the time to death are recorded.

We must first set up the data so that they are in the format required to use the stpm2 and stpm2illd commands.

```
use rott2
(Rotterdam breast cancer data, truncated at 10 years)
. illdprep, id(pid) statevar(rf osi) statetime(rf os) addtime(0.1)
Note that .1 has been added to os for one or more individuals as the addtime option has been specified by the user. These individuals are indicated with a value of 1 in the newly generated _check variable.
Note that one or more individuals have the rfi event at the same time as they are censored for the rfi event. The program assumes that the individual was not at risk of osi after the rfi time and therefore will not have a third row in the data. These individuals are indicated with a value of 1 in the newly generated _check2 variable. The user may wish to change this in the original data and rerun the command.
The data have been expanded so that each patient has up to three rows of data as demonstrated in tables 1 and 2. Three indicator variables have been created for each of the three transitions (trans1, trans2, and trans3). A variable, trans, is also stored in the data and will be needed to obtain initial values in the stpm2 command. A further indicator variable called status has been created to summarize which of the three transitions each patient has experienced: 1 indicates that the patient has experienced the transition, and 0 indicates otherwise. The addtime() option has been specified to add 0.1 to the death time for any patients who relapse and die at the exact same time. The relapse and death times are in months from diagnosis; thus 0.1 is equivalent to approximately 3 days in this example. A _check variable has been generated in correspondence with 0.1 to indicate which patients had this amount added to their death time. A warning has also been given for one or more patients who have a relapse and are censored for the death event at the same time. This means that for such a patient, the command has dropped the third row of data representing transition 3 because the patient was never actually at risk of death after relapse. Finally, the command has generated start and stop times to show when a patient enters and exits each state. These newly generated variables can be used to stset the data. We can then run the stpm2 command for all three transitions simultaneously.
```
S. R. Hinchliffe, D. A. Scott, and P. C. Lambert

```
. stset stop, enter(start) failure(status==1) scale(12)
  failure event:  status == 1
  obs. time interval:  (0, stop]
  enter on or after:  time start
  exit on or before:  failure
  t for analysis:  time/12

7471  total obs.
  0  exclusions

7471  obs. remaining, representing
2790  failures in single record/single failure data
38398.57  total analysis time at risk, at risk from t = 0
  earliest observed entry t = 0
  last observed exit t = 19.28268

. stpm2 trans1 trans2 trans3 age, scale(hazard) rcsbaseoff nocons dftvc(3)
> tvc(trans1 trans2 trans3) initstrata(trans) eform

note: delayed entry models are being fitted

Iteration 0:  log likelihood =  -5497.7319
Iteration 1:  log likelihood =  -5495.6716
Iteration 2:  log likelihood =  -5495.6418
Iteration 3:  log likelihood =  -5495.6418

Log likelihood =  -5495.6418  Number of obs    =  7471

|          | exp(b)  | Std. Err. | z   | P>|z|  |       95% Conf. Interval       |
|----------|---------|-----------|-----|------|--------------------------------|
|          |         |           |     |      |                                |
| xb       |         |           |     |      |                                |
| trans1   | .02331  | .0028974  | -30.24 | 0.000 | .01827 - .0297403             |
| trans2   | .2455235 | .0216091 | -11.56 | 0.000 | .206622 - .291749             |
| trans3   | .9442842 | .1211267 | -0.45 | 0.655 | .7343719 - 1.214198           |
| age      | 1.008449 | .0015035 | 5.64 | 0.000 | 1.005507 - 1.0114             |
| _rcs_trans11 | 3.537942 | .3075088 | 14.54 | 0.000 | 2.983778 - 4.195029           |
| _rcs_trans12 | .9383132 | .0507433 | -1.18 | 0.239 | .8439475 - 1.04323            |
| _rcs_trans13 | .9906213 | .0352729 | -0.63 | 0.527 | .9238449 - 1.057394           |
| _rcs_trans21 | 2.539793 | .0574909 | 41.18 | 0.000 | 2.429576 - 2.65501            |
| _rcs_trans22 | 1.29505 | .0241911 | 13.84 | 0.000 | 1.248949 - 1.341342           |
| _rcs_trans23 | .9669232 | .0094508 | -3.44 | 0.001 | .9488572 - .985625           |
| _rcs_trans31 | 2.171531 | .209309 | 8.50 | 0.000 | 1.797714 - 2.62308           |
| _rcs_trans32 | 1.162727 | .0698784 | 2.51 | 0.012 | 1.033527 - 1.308079           |
| _rcs_trans33 | .9826401 | .0147 | -1.17 | 0.242 | .9542469 - 1.011878          |
```
Patients can be at risk of death with relapse only after they have experienced the relapse event; therefore, the time for this state is later than the time of origin. This means that a delayed entry model is fit as indicated in the \texttt{stpm2} command. By default, the \texttt{stpm2} command obtains initial values from a Cox model. The \texttt{initstrata()} option in the command line allows for this Cox model to be stratified by the three transitions. By including the three transition indicators (\texttt{trans1()}, \texttt{trans2()}, and \texttt{trans3()}) as both main effects and time-dependent effects (using the \texttt{tvc()} option), we have fit a stratified model with three separate baselines, one for each transition. For this reason, we have used the \texttt{rcsbaseoff} option together with the \texttt{nocons} option, which excludes the baseline hazard from the model. The hazard ratio (95\% confidence intervals) for age is 1.008449 (1.005507 to 1.0114). This means that all three transition rates increase by 0.8\% with each yearly increase in age. By including age in the model in this way, we have assumed that the effect of age remains constant across all three transitions. This is unlikely to be the case.

By including interaction terms between age and the three transition indicators, we can estimate a different age effect for each transition.

\begin{verbatim}
. forvalues i=1/3 {
2.    generate trans'\_iage'=trans'\_i'*age
3. }
. stpm2 trans1 trans2 trans3 trans1age trans2age trans3age,
   > scale(hazard) rcsbaseoff nocons dftvc(2)
   > tvc(trans1 trans2 trans3) initstrata(trans) eform
note: delayed entry models are being fitted
Iteration 0: log likelihood = -5369.4658
Iteration 1: log likelihood = -5332.4523
Iteration 2: log likelihood = -5330.8393
Iteration 3: log likelihood = -5330.8192
Iteration 4: log likelihood = -5330.8191
Log likelihood = -5330.8191 Number of obs = 7471

| exp(b) | Std. Err. | z     | P>|z|   | [95\% Conf. Interval] |
|--------|-----------|-------|-------|-------------------|
| xb     |           |       |       |                   |
| trans1 | 8.91e-06  | 5.07e-06 | -20.41 | 0.000          | 2.92e-06      | 0.0000272 |
| trans2 | 0.4515908 | 0.0521228 | -8.89  | 0.000          | 0.3601785     | 0.5662032 |
| trans3 | 1.181057  | 0.1969131 | 1.00   | 0.318          | 0.8518095     | 1.637527 |
| trans1age | 1.139042 | 0.089574  | 16.55  | 0.000          | 1.121621      | 1.156735  |
| trans2age | 0.9974217 | 0.0020578 | -1.25  | 0.211          | 0.9839666     | 1.001463  |
| trans3age | 1.006303 | 0.0023563 | 2.68   | 0.007          | 1.001696      | 1.010932  |
| \_rcs_trans11 | 3.951158 | 0.338209 | 16.02  | 0.000          | 3.339888      | 4.674303  |
| \_rcs_trans12 | 0.8822663 | 0.0454067 | -2.43  | 0.015          | 0.7976121     | 0.9759051 |
| \_rcs_trans21 | 2.493473 | 0.043812 | 41.89  | 0.000          | 2.389134      | 2.602369  |
| \_rcs_trans22 | 1.240989 | 0.0179256 | 14.95  | 0.000          | 1.206348      | 1.276624  |
| \_rcs_trans31 | 1.939866 | 0.1551909 | 8.28   | 0.000          | 1.658365      | 2.269198  |
| \_rcs_trans32 | 1.078697 | 0.035531 | 2.30   | 0.021          | 1.011258      | 1.150633  |
\end{verbatim}

The hazard ratio (95\% confidence interval) for the age transition 1 interaction is 1.139042 (1.121621 to 1.156735), which suggests that the transition rate from alive to dead increases by approximately 14\% with every yearly increase in age. The hazard ratio (95\% confidence interval) for the age transition 2 interaction is 0.9974217 (0.9933966 to
1.001463), which suggests that the transition rate from alive to relapse decreases with age; however, this is not significant. Finally, the hazard ratio (95% confidence interval) for the age transition 3 interaction is 1.006303 (1.001696 to 1.010932), which suggests that for those who relapse, the transition rate from relapse to dead also increases with age.

Now that we have run stpm2, we can run the postestimation command stpm2illd to obtain the probability of being in each of the four states as demonstrated in figure [1]. Because we have included age as a continuous variable, we need to choose a particular covariate pattern for which to make the predictions. We will run the stpm2illd command twice, once for age 65 and once for age 85.

```
. * Age 65 *
. stpm2illd alive65 relapse65 death65 relapsedeath65,
  > trans1(trans1 1 trans1age 65) trans2(trans2 1 trans2age 65)
  > trans3(trans3 1 trans3age 65) ci
. * Age 85 *
. stpm2illd alive85 relapse85 death85 relapsedeath85,
  > trans1(trans1 1 trans1age 85) trans2(trans2 1 trans2age 85)
  > trans3(trans3 1 trans3age 85) ci
```

The `trans1()` to `trans3()` options give the linear predictor for each of the three transitions for which we want the prediction. The commands have generated eight new variables containing the probabilities of being in each state. The predictions for age 65 are denoted with a 65 at the end of the variable name, and the predictions for age 85 are denoted with an 85. The eight probabilities are `prob_alive65`, `prob_ill65`, `prob_death65`, `prob_illdeath65`, `prob_alive85`, `prob_ill85`, `prob_death85`, and `prob_illdeath85`. Each of these variables has a corresponding high and low confidence bound, for example, `prob_alive65_lci` and `prob_alive65_uci`. These were created when the `ci` option was specified.
If we plot the probability of each state along with its confidence intervals against time for both age 65 and age 85, we can achieve plots as shown in figure 2.

Figure 2 shows that the probability of remaining alive and well is significantly lower for those aged 85 compared with those aged 65. By 15 years, the probability of being alive and well is almost 0 for those aged 85. As expected, the probability of dying before relapse is higher for those aged 85, with values reaching approximately 0.63 by 15 years compared with 0.15 for those aged 65.
The plot for the probability of relapse is different in shape from the other three plots. This is because relapse is a transient state; patients may enter the relapse state, but after some time, they may leave that state and go on to die. This gives the curve that peaks after about 3 or 4 years for both those aged 65 (probability approximately 0.2) and those aged 85 (probability approximately 0.18). The curve then begins to decrease as more patients with relapse go on to die. Finally, the probability of death for those that suffer a relapse is higher at age 65 (approximately 0.48) than at age 85 (approximately 0.34). This is due to the high number of deaths before relapse in those aged 85.

The model shown above assumes proportional hazards for the age transition interactions. In many epidemiological studies, the effect of age will be time dependent. We will now fit the flexible parametric survival model again and include time-dependent effects for the age transition interactions. This time, we want to obtain only one estimate for the overall probability of death, that is, to combine the probabilities of being in stages 3 and 4 in figure 1. To do this, we need to use the combine option. When we use this option, we only need to specify three new variable names in the stpm2illd command line.
Flexible parametric illness-death models

```
.s SMP2 trans1 trans2 trans3 trans1age trans2age trans3age, scale(hazard) rcsbaseoff nocons
> dftvc(trans1age:2 trans2age:2 trans3age:2 3)
> tvc(trans1 trans2 trans3 trans1age trans2age trans3age)
> initstrata(trans) eform

note: delayed entry models are being fitted
```

Iteration 0: log likelihood = -5324.6353
Iteration 1: log likelihood = -5311.9706
Iteration 2: log likelihood = -5310.9136
Iteration 3: log likelihood = -5310.8708
Iteration 4: log likelihood = -5310.8707

Log likelihood = -5310.8707 Number of obs = 7471

<table>
<thead>
<tr>
<th>exp(b)</th>
<th>Std. Err.</th>
<th>z</th>
<th>P&gt;z</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>xb</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trans1</td>
<td>.00001</td>
<td>6.98e-06</td>
<td>-16.52</td>
<td>0.000</td>
</tr>
<tr>
<td>trans2</td>
<td>.4132457</td>
<td>.0503478</td>
<td>-7.25</td>
<td>0.000</td>
</tr>
<tr>
<td>trans3</td>
<td>.7806882</td>
<td>.2194512</td>
<td>-3.57</td>
<td>0.000</td>
</tr>
<tr>
<td>trans1age</td>
<td>1.137403</td>
<td>.0109405</td>
<td>13.38</td>
<td>0.000</td>
</tr>
<tr>
<td>trans2age</td>
<td>.9989598</td>
<td>.002172</td>
<td>-0.48</td>
<td>0.632</td>
</tr>
<tr>
<td>trans3age</td>
<td>1.011249</td>
<td>.0048782</td>
<td>-0.48</td>
<td>0.632</td>
</tr>
<tr>
<td>_rcs_trans11</td>
<td>4.143021</td>
<td>2.836773</td>
<td>2.08</td>
<td>0.038</td>
</tr>
<tr>
<td>_rcs_trans12</td>
<td>1.55668</td>
<td>.7534786</td>
<td>-0.41</td>
<td>0.685</td>
</tr>
<tr>
<td>_rcs_trans13</td>
<td>.9768487</td>
<td>.0402173</td>
<td>-0.57</td>
<td>0.569</td>
</tr>
<tr>
<td>_rcs_trans21</td>
<td>3.084326</td>
<td>.2939969</td>
<td>10.78</td>
<td>0.000</td>
</tr>
<tr>
<td>_rcs_trans22</td>
<td>1.552191</td>
<td>.1146111</td>
<td>6.12</td>
<td>0.000</td>
</tr>
<tr>
<td>_rcs_trans23</td>
<td>.9740596</td>
<td>.0096225</td>
<td>-2.66</td>
<td>0.008</td>
</tr>
<tr>
<td>_rcs_trans31</td>
<td>3.232405</td>
<td>.8243798</td>
<td>-4.00</td>
<td>0.000</td>
</tr>
<tr>
<td>_rcs_trans32</td>
<td>1.59504</td>
<td>.2284165</td>
<td>3.28</td>
<td>0.001</td>
</tr>
<tr>
<td>_rcs_trans33</td>
<td>.987701</td>
<td>.0144313</td>
<td>-0.85</td>
<td>0.397</td>
</tr>
<tr>
<td>_rcs_trans1age1</td>
<td>.9992748</td>
<td>.0053111</td>
<td>-0.08</td>
<td>0.938</td>
</tr>
<tr>
<td>_rcs_trans1age2</td>
<td>.9922872</td>
<td>.0064791</td>
<td>-1.19</td>
<td>0.236</td>
</tr>
<tr>
<td>_rcs_trans2age1</td>
<td>.9965224</td>
<td>.0016427</td>
<td>-2.11</td>
<td>0.035</td>
</tr>
<tr>
<td>_rcs_trans2age2</td>
<td>.99681</td>
<td>.001214</td>
<td>-2.62</td>
<td>0.009</td>
</tr>
<tr>
<td>_rcs_trans3age1</td>
<td>.9937488</td>
<td>.0039773</td>
<td>-1.57</td>
<td>0.117</td>
</tr>
<tr>
<td>_rcs_trans3age2</td>
<td>.994577</td>
<td>.0020061</td>
<td>-2.51</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Notice that we have allowed different degrees of freedom for the age transition interactions (2df) and the three separate transition baselines (3df) by specifying this in the `dftvc()` option. We have also dropped the variables generated in the previous `stpm2illd` command. If users did not wish to do this, then they would have to specify different names for the probability variables when running the command again. Rather than graphing the probabilities of being in each state as separate line plots (as we did previously), we can display them by stacking the probabilities on top of one another. This produces a graph as shown in figure 3. To do this, we need to generate new vari-
ables that sum up the probabilities. This is done for each of the two age predictions, 65 and 85. The code shown below is for those aged 85 only.

```stata
. generate tot1=prob_alive85
  (6471 missing values generated)
. generate tot2=prob_alive85+prob_relapse85
  (6471 missing values generated)
. generate tot3=prob_alive85+prob_relapse85+prob_death85
  (6471 missing values generated)
. twoway (area tot3 _newt if _newt<=15, sort)
  > (area tot2 _newt if _newt<=15, sort) (area tot1 _newt if _newt<=15, sort),
  > legend(order(3 "Alive and well" 2 "Relapse" 1 "Dead") rows(1))
  > ylabel(0(0.2)1, angle(0) format(%3.1f))
  > xlabel("Time since diagnosis (years)") xtitle("Age 85")
  > plotregion(margin(zero)) scheme(sj)
```

![Figure 3. Stacked probability of being alive, having a relapse, and dying as a function of time since diagnosis (years) for those aged 65 and 85](image)

Figure 3. Stacked probability of being alive, having a relapse, and dying as a function of time since diagnosis (years) for those aged 65 and 85
As we showed previously in figure 2, the probability of remaining alive and well for those aged 85 decreases to almost 0 over the period of 15 years. The probability of being alive after relapse is highest between approximately 1 and 5 years since breast cancer diagnosis for those aged 85. It then starts to decrease as more patients die with relapse. For those aged 65, the probability of being alive after relapse remains fairly stable beyond 5 years. By 15 years, approximately 65% of those aged 65 and 98% of those aged 85 have died.

6 Conclusion

The new commands `illdprep` and `stpm2illd`, in conjunction with the existing command `stpm2`, provide a suite of programs that will enable users to estimate transition hazards and probabilities within an illness-death model framework using flexible parametric survival models. We hope that it will be a useful tool in medical research. The illness-death model is a very simple multistate model. Therefore, further developments are needed to fit more complex multistate models.

7 References


**About the authors**

Sally Hinchliffe is a PhD student at the University of Leicester, UK. She is currently working on developing methodology for application in competing risks.

David Scott is Senior Director of Health Economics at Oxford Outcomes Ltd and an MSc student in Medical Statistics at the University of Leicester, UK, where he is undertaking his thesis on multistate modeling.

Paul Lambert is a professor of biostatistics at the University of Leicester, UK. His main interest is in the development and application of methods in population-based cancer research.