



*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](mailto: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.*

**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. SCHAFER, 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>

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

U.S. and Canada		Elsewhere	
<b>Printed &amp; electronic</b>		<b>Printed &amp; electronic</b>	
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
<b>Electronic only</b>		<b>Electronic only</b>	
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](mailto: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.

# Frailty models and frailty-mixture models for recurrent event times

Ying Xu

Center for Quantitative Medicine  
Duke–NUS Graduate Medical School  
Singapore

and Department of Biostatistics  
Singapore Clinical Research Institute  
Singapore

tina.xuying@duke-nus.edu.sg

Yin Bun Cheung

Center for Quantitative Medicine  
Duke–NUS Graduate Medical School  
Singapore

and Department of International Health  
University of Tampere  
Finland

yinbun.cheung@duke-nus.edu.sg

**Abstract.** The analysis of recurrent event times faces three challenges: between-subject heterogeneity (frailty), within-subject event dependence, and the possibility of a cured fraction. Frailty can be handled by including a latent random-effects term in a Cox-type model. Event dependence may be considered as contributing to the intervention effect, or it may be considered as a source of nuisance, depending on the analysts’ specific research questions. If it is seen as a nuisance, the analysis can stratify the recurrent event times according to event order. If it is seen as contributing to the intervention effect, stratification should not be used. Models with and without stratification for event order estimate two types of treatment effects. They are analogous to per-protocol analysis and intention-to-treat analysis, respectively. In the context of chronic disease treatment, we want to estimate whether there is a cured fraction; for infectious disease prevention, this is called a nonsusceptible fraction. In infectious disease prevention, we want to understand whether an intervention protects each of its recipients to some extent (“leaky” model) or whether it totally protects some recipients but offers no protection to the rest (“all-or-none” model). The truth may be a mixture of the two modes of protection. We describe a class of regression models that can handle all three issues in the analysis of recurrent event times. The model parameters are estimated by the expectation-maximization algorithm, and their variances are estimated by Louis’s formula. We provide a new command, `strmcure`, for implementing these models.

**Keywords:** st0374, strmcure, frailty models, frailty-mixture models, recurrent event times, event dependence, cured fraction

## 1 Introduction

Recurrent event times are common in biomedical and social studies. Some examples include times to respiratory symptom exacerbations, hospital readmissions, and malaria disease episodes. Compared with the analysis of time-to-first or only event, the analysis of recurrent event times offers some advantages. For example, in analysis of time-to-first event, the unobserved heterogeneity (or frailty) impacts on the individual follow-up time (or person-time): subjects who are more frail will experience their first events and

become censored earlier than subjects who are less frail. This causes biased results such as an attenuated estimate for the effect of intervention (or exposure). Greenland (1996) gives an example of studying the exposure effect on first headache, in which ignoring the effect of genotype on the person-time would lead to distorted estimates for the exposure effect. However, there is no such problem in the analysis of recurrent event times because the observations are not censored by the first event; therefore, heterogeneity does not affect the person-time. This would be the case in the example of headaches “when the study outcome had been ‘headache’, which can recur, rather than ‘first headache’” (Greenland 1996, 500). Therefore, using recurrent event times when evaluating interventions can provide a better understanding of the intervention effects. The World Health Organization Malaria Vaccine Advisory Committee has recently decided to use recurrent event times to evaluate malaria vaccines and also for further methodological research on how best to analyze recurrent event times (Moorthy, Reed, and Smith 2009).

Analysis of recurrent event times faces three challenges. First, some frailty is always present because of omitted or unobserved covariates (Aalen 1988, 1121). “Substantial frailty may therefore be more common than generally supposed” (Pickles and Crouchley 1994, 264) because of the typical use of specification tests that are not sufficiently sensitive to the presence of frailty. Frailty causes a bias not only in the estimate of the level of hazards but also in the hazard ratio. In the analysis of a single event, without replications on the subject level, there is limited power to distinguish the random variation as within subject (large spread of baseline hazard function over time) or between subject (large spread of frailty distribution), which often leads to unstable behavior in the model estimation (Keiding, Anderson, and Klein 1997). In contrast, incorporating a latent random-effects term in models for recurrent event times is possible and is robust to misspecification of the frailty distribution (for example, Pickles and Crouchley [1995]; O’Quigley and Stare [2002]; Xu et al. [2012]).

Second, the presence of within-subject event dependence complicates the analysis. This is different from frailty. Event dependence refers to an event occurrence changing the hazard in the future. For example, an episode of myocardial infarction damages the heart muscle. This damage then increases the chance that the person may have more episodes of myocardial infarction (Metcalf and Thompson 2006). We call this positive event dependence. Negative event dependence is also possible. For example, people who have experienced injuries may learn from the events. The behavioral changes lead to a lower chance of future injuries (Ullah, Gabbett, and Finch 2014). If there is positive event dependence, the success in preventing one event has a secondary effect on reducing the rate of subsequent events. If there is negative event dependence, a short-term success may be neutralized by a relatively sustained event rate in the long term. As such, in the evaluation of an intervention, the total effect depends on a primary effect and a secondary effect, where the secondary effect is via event dependence (Cheung et al. 2010; Simpson 2013). The estimates of the primary effect and total effect are analogous to per-protocol analysis and intention-to-treat analysis, respectively. From a public health point of view, the total effect is more important. But from a product or program development point of view, the primary effect is also important (Cheung et al.

2010). Some statistical models (for example, the Prentice–Williams–Peterson model [Prentice, Williams, and Peterson 1981] and the conditional frailty model) stratify the data according to the event order. (See Kelly and Lim [2000], Therneau and Grambsch [2000], and Box-Steffensmeier and De Boef [2006] for overviews of such models.) By using event-order stratification, the effect of event dependence is incorporated in the unspecified event-specific baseline hazards. Therefore, they estimate the primary effect. In contrast, models without stratification allow an event to affect the later event rate. They estimate the total effect. The choice between using a model with or without stratification depends on what one wants to estimate (the primary or total effects). It is important to choose a model on the premise of the specific research question. This usually requires collaboration between subject-matter experts and statisticians.

Third, there has long been interest in whether a fraction of the population is event free even after long-term follow-up. In the context of chronic disease treatment, this is called a cured fraction. In the context of infectious disease prevention, this is called a nonsusceptible fraction. In infectious disease prevention, there is interest in whether an intervention protects each of its recipients to some extent (“leaky” model) or whether it totally protects some recipients but offers no protection to the rest (“all-or-none” model) (Smith, Rodrigues, and Fine 1984; Halloran, Longini, and Struchiner 1996). The truth may be a mixture of the two modes of protection.

In this article, we describe frailty-mixture models for regression analysis of recurrent event times recently proposed by Xu et al. (2012). We also describe a new command, **strmcure**, for fitting frailty models for recurrent event times with or without stratification for event order and with or without a cured fraction. The user controls the choice of time scale (counting process or gap times) with **stset**, which further enriches the range of models that can be fitted (Kelly and Lim 2000; Therneau and Grambsch 2000). The article is structured as follows: In section 2, we present the frailty-mixture models. In section 3, we describe the estimation approach. In section 4, we describe the syntax of **strmcure** and explain its options. In section 5, we illustrate the command using a clinical trial dataset, and we discuss the macro functions in relation to **stcox**. We conclude in section 6.

## 2 Model specifications

Suppose we have a random sample of  $n$  subjects. We use subscripts  $i$  and  $j$  to index the subject and the sequence of the recurrent event times within the same subject, respectively. For subject  $i$ , let  $t_{ij}$  denote the  $j$ th event time ( $j = 1, \dots, n_i$ ) and  $\delta_{ij}$  denote the event indicator, which takes the value of 0 if the  $j$ th event time is right-censored and 1 otherwise. Let  $d_i$  denote the number of observed events in subject  $i$ ; that is,  $d_i = \sum_{j=1}^{n_i} \delta_{ij}$ . Event time can be gap time or counting process time. Gap time is the time from study enrollment for  $j = 1$  and time since the previous event for  $j > 1$ . That is, the clock resets to zero after each event. Counting process does not reset the clock.

We define a latent binary variable  $k_i$ :  $k_i = 0$  if subject  $i$  is cured, and  $k_i = 1$  if he or she is noncured and will eventually experience the event. The latent variable  $k_i = 1$  can be fixed at 1 for all subjects to indicate that the model does not include a cured fraction or that it can be modeled by

$$\pi_i = \Pr(k_i = 1 | \mathbf{Z}_i) = g(\mathbf{Z}_i \boldsymbol{\theta}) \quad (1)$$

where  $g(\cdot)$  is a link function,  $\mathbf{Z}_i$  is a vector of baseline covariates, and  $\boldsymbol{\theta}$  is the regression coefficient vector. The link functions include the logit, probit, and complementary log-log link.

If subject  $i$  is noncured ( $k_i = 1$ ), the hazard function at the  $j$ th event time  $t$  is, as Xu et al. (2012) show,

$$h_{ij}(t | Y_{ij}(t), \omega_i, \mathbf{X}_{ij}(t), k_i = 1) = Y_{ij}(t) \omega_i h_{0j}(t) \exp \{ \mathbf{X}_{ij}^T(t) \boldsymbol{\beta} \} \quad (2)$$

where  $Y_{ij}(t)$  is the at-risk indicator:  $Y_{ij}(t) = 1$  if the subject is at risk for the  $j$ th event at time  $t$ , and  $Y_{ij}(t) = 0$  otherwise. Given that  $k_i = 1$ ,  $\omega_i$  is the subject-specific frailty and is assumed to follow a  $\Gamma(\psi^{-1}, \psi)$  distribution where  $\psi$  is the variance of the frailty distribution. Mathematically, the probability density function for  $\omega_i$  is  $f(\omega_i | k_i = 1) = \omega_i^{1/\psi-1} \exp(-\omega_i/\psi) / \Gamma(1/\psi) \psi^{1/\psi}$ . Moreover,  $h_{0j}(\cdot)$  is the unspecified baseline hazard function specific to the  $j$ th event. It reduces to  $h_0(\cdot)$  if there is no stratification for event order.  $\mathbf{X}_{ij}(t)$  is a vector of possibly time-varying covariates, and  $\boldsymbol{\beta}$  is the regression coefficient vector.

The  $\pi_i$  and  $h_{ij}(\cdot)$  parts in (1) and (2) together define our estimation model. Let  $\boldsymbol{\Omega}_i = \{ \mathbf{X}_{ij}(t), Y_{ij}(t), \mathbf{Z}_i, t_{ij}, \delta_{ij} : 0 < t \leq t_{ij}, j = 1, \dots, n_i \}$  denote the observed data in subject  $i$ , and let  $\boldsymbol{\Omega} = (\boldsymbol{\Omega}_i : i = 1, \dots, n)$  denote the observed data in the study sample. The complete data likelihood for subject  $i$  (denoted by  $L_{i,c}$ , conditional on  $\boldsymbol{\Omega}_i$ ,  $k_i$ , and  $\omega_i$ ) can be written as

$$L_{i,c} = (1 - \pi_i)^{1-k_i} (\pi_i G_i)^{k_i} \quad (3)$$

where

$$G_i = \frac{\omega_i^{1/\psi-1} \exp(-\omega_i/\psi)}{\Gamma(1/\psi) \psi^{1/\psi}} \prod_{j=1}^{n_i} \left( [\omega_i h_{0j}(t_{ij}) \exp \{ \mathbf{X}_{ij}^T(t_{ij}) \boldsymbol{\beta} \}]^{\delta_{ij}} \exp \{ -\omega_i H_{ij}(t_{ij}) \} \right)$$

and

$$H_{ij}(t_{ij}) = \int_0^{t_{ij}} Y_{ij}(s) h_{0j}(s) \exp \{ \mathbf{X}_{ij}^T(s) \boldsymbol{\beta} \} ds$$

The complete data likelihood of all the study subjects, conditional on  $\boldsymbol{\Omega}$ ,  $k_i$ , and  $\omega_i$  ( $i = 1, \dots, n$ ), is  $L_c = \prod_{i=1}^n L_{i,c}$ .

### 3 Estimation approach

The parameter of interest is denoted by  $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \psi)^T$ . Furthermore, we define  $\boldsymbol{\zeta} = (\boldsymbol{\eta}^T, \mathbf{h}_0^T)^T$ , where  $\mathbf{h}_0 = (\mathbf{h}_{01}, \dots, \mathbf{h}_{0J})^T$  and  $\mathbf{h}_{0j}$ ,  $j = 1$  to  $J$ , as a column vector with its elements being the baseline hazards at all the uncensored  $j$ th event times observed in the study sample.

The complete data likelihood from (3) involves unobserved variables  $k_i$  and  $\omega_i$ . One approach of parameter estimation is a gradient-based maximum-likelihood estimation technique. This requires one to calculate the observed data-likelihood contribution from each subject by integrating the complete data-likelihood contribution in (3) over the unknown variables  $k_i$  and  $\omega_i$ . The parameters in  $\zeta$  can be estimated simultaneously. Nevertheless, the implementation of this maximum-likelihood estimation approach is nontrivial or even infeasible sometimes. This is because the dimension of the parameter space and the dimension of the associated Hessian matrix depend on the number of distinct event times observed in the study sample. When this number is large, the creating and inverting of a high-dimensional Hessian matrix could be challenging and easily exhaust system memory. Moreover, the numerical inversion of a high-dimension Hessian matrix may be difficult or inaccurate at some iteration. To circumvent these difficulties, we consider an alternative estimation algorithm—the expectation-maximization (EM) algorithm—for parameter estimation. The EM algorithm is a convenient tool for parameter estimation and treats the unobserved variables as missing. It is an iterative process for parameter estimation, with each iteration involving an expectation step followed by a maximization step. The methodology underlying the EM algorithm is in Dempster, Laird, and Rubin (1977). The implementation of the EM algorithm is summarized as follows (more detailed information can be found in Xu et al. [2012]):

**Step 0.** We provide the initial parameter estimates for  $\theta, \beta, \psi$ , and  $h_0$ . We arbitrarily partition the study sample into a cured group (consisting of subjects whose first event is right-censored) and a noncured group (consisting of subjects whose first event is observed). Using this partition, we fit a binary regression model (with the existing Stata command `logistic` or `probit` or `cloglog`, depending on the user-specified link function) to obtain the initial estimate  $\hat{\theta}^{(0)}$ . Then, we fit a standard Cox model, using the `stcox` command, to the observed event data on subjects in the noncured group to obtain the initial estimates,  $\hat{\beta}^{(0)}$  and  $\hat{h}_0^{(0)}$ . The initial estimate for  $\psi$  is arbitrarily assigned as  $\hat{\psi}^{(0)} = 2$  (or specified by the user with the `phi0()` option in the `strmcure` command).

**Step 1** (expectation step). In the  $r$ th iteration (for  $r = 1, \dots$ ), we compute the expectations of  $k_i$ ,  $k_i \omega_i$ , and  $k_i \log \omega_i$  ( $i = 1, \dots, n$ ). On the basis of (3), the conditional probability of being noncured ( $k_i = 1$ ) is

$$\Pr(k_i = 1) = \begin{cases} 1 & \text{if } d_i > 0 \\ \frac{\pi_i \{1 + \psi H_{i1}(t_{i1})\}^{\frac{-1}{\psi}}}{1 - \pi_i + \pi_i \{1 + \psi H_{i1}(t_{i1})\}^{\frac{-1}{\psi}}} & \text{if } d_i = 0 \end{cases} \quad (4)$$



The conditional distribution of the frailty  $\omega_i$  if  $k_i = 1$  is

$$\omega_i | (k_i = 1) \sim \text{Gamma}(\psi^{-1} + d_i, \varrho_i^{-1}) \quad (5)$$

with  $\varrho_i = \psi^{-1} + \sum_{j=1}^{m_i} H_{ij}(t_{ij})$ . If  $k_i = 0$ , then  $\omega_i$  is undefined. On the basis of (4) and (5), it follows that

$$\begin{aligned} E(k_i) &= \Pr(k_i = 1) \\ E(k_i \omega_i) &= \frac{E(k_i)(\psi^{-1} + d_i)}{\varrho_i} \\ E(k_i \log \omega_i) &= E(k_i) \{ \text{digamma}(\psi^{-1} + d_i) - \log \varrho_i \} \end{aligned}$$

where digamma is the `digamma()` function. The expectations of  $k_i$ ,  $k_i \omega_i$ , and  $k_i \log \omega_i$  can then be estimated given the observed data and the parameter estimates  $\hat{\boldsymbol{\theta}}^{(r-1)}$ ,  $\hat{\boldsymbol{\beta}}^{(r-1)}$ ,  $\hat{\psi}^{(r-1)}$ , and  $\hat{\mathbf{h}}_0^{(r-1)}$  from the  $(r-1)$ th iteration.

**Step 2** (maximization step). Conditional on the estimated expectations of  $k_i$ ,  $k_i \omega_i$ , and  $k_i \log \omega_i$  from step 1, we then update the parameter estimates for  $\boldsymbol{\eta} = (\boldsymbol{\theta}^T, \boldsymbol{\beta}^T, \psi)^T$  by solving the estimating equation  $U_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = 0$  derived from (3). Here  $U_{\boldsymbol{\eta}}(\boldsymbol{\eta}) = \{U_{\boldsymbol{\theta}}^T(\boldsymbol{\theta}), U_{\boldsymbol{\beta}}^T(\boldsymbol{\beta}), U_{\psi}(\psi)\}^T$ , in which

$$U_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \sum_{i=1}^n \left( \frac{k_i}{\pi_i} - \frac{1 - k_i}{1 - \pi_i} \right) g'(\mathbf{Z}_i^T \boldsymbol{\theta}) \mathbf{Z}_i \quad (6)$$

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = \sum_{i=1}^n \sum_{j=1}^{n_i} \delta_{ij} \left[ \mathbf{X}_{ij}(t_{ij}) - \frac{\sum_{l=1}^n Y_{lj}(t_{ij}) k_l \omega_l \exp\{\mathbf{X}_{lj}^T(t_{ij}) \boldsymbol{\beta}\} \mathbf{X}_{lj}(t_{ij})}{\sum_{l=1}^n Y_{lj}(t_{ij}) k_l \omega_l \exp\{\mathbf{X}_{lj}^T(t_{ij}) \boldsymbol{\beta}\}} \right] \quad (7)$$

$$U_{\psi}(\psi) = \frac{1}{\psi^2} \sum_{i=1}^n [k_i \omega_i - k_i \log \omega_i + k_i \log \{\Gamma(\psi^{-1})\} + k_i \log \psi - k_i] \quad (8)$$

The scalar function  $g'(w)$  in (6) is the first derivative of the link function  $g(w)$  with respect to  $w$ . We solve (6) to (8) with `strmcure_mylogit_d2` (or `strmcure_myprobit_d2` or `strmcure_mycloglog_d2`, depending on the user-specified link function for the binary regression), `stcox`, and `strmcure_myGamma_d2`, respectively.

The estimate for the baseline hazard,  $h_{0j}(t_{ij})$ , is then updated using the Nelson–Aalen estimator as follows:

$$h_{0j}(t_{ij}) = \frac{\delta_{ij}}{\sum_{l=1}^n Y_{lj}(t_{ij}) k_l \omega_l \exp\{\mathbf{X}_{lj}^T(t_{ij}) \boldsymbol{\beta}\}} \quad (9)$$

Because data might become sparse at higher-event strata, the last few strata may be pooled. We denote the parameter estimates obtained at this step as  $\hat{\boldsymbol{\theta}}^{(r)}$ ,  $\hat{\boldsymbol{\beta}}^{(r)}$ ,  $\hat{\psi}^{(r)}$ , and  $\hat{\mathbf{h}}_0^{(r)}$ .

**Step 3.** We iterate between steps 1 and 2 until the convergence criterion is met. We denote the estimates at convergence by  $\hat{\zeta}$ .

If the model does not involve a cured fraction, then we have  $k_i = 1$  ( $i = 1, \dots, n$ ) for all subjects, and we ignore the estimation part for the parameter  $\theta$  when implementing the above EM algorithm.

We estimate the variance–covariance matrix of the parameter estimates by inverting the observed Fisher information matrix (or the Fisher information matrix for the observed data likelihood). The EM algorithm for the incomplete-data problems provides only point estimates for the parameters; it does not provide the observed Fisher information matrix directly. Louis (1982) provided a formula for calculating the Fisher information matrix within the EM context. Louis’s formula requires the user to compute the first and second derivatives of the complete data log likelihood, and these need be computed only at the last iteration of the EM algorithm. Louis’s formula for obtaining the Fisher information matrix is

$$I(\zeta) = -E \left( \frac{\partial^2 \log L_c}{\partial \zeta \partial \zeta^T} \middle| \zeta, \Omega \right) - E \left( \frac{\partial \log L_c}{\partial \zeta} \frac{\partial \log L_c}{\partial \zeta^T} \middle| \zeta, \Omega \right) + E \left( \frac{\partial \log L_c}{\partial \zeta} \middle| \zeta, \Omega \right) E \left( \frac{\partial \log L_c}{\partial \zeta^T} \middle| \zeta, \Omega \right) \quad (10)$$

where  $\partial \log L_c / \partial \zeta$  and  $\partial^2 \log L_c / \partial \zeta \partial \zeta^T$  are, respectively, the first and second derivatives of the complete-data log-likelihood  $\log L_c$  with respect to the parameter vector,  $\zeta$ . The expectations in (10) are all taken over the unobserved  $k_i$  and  $\omega_i$  ( $i = 1, \dots, n$ ).

Theoretical derivations of the expectations are usually daunting and not always possible, such as in our context of a frailty-mixture model with unspecified baseline hazard function. In practice, the expectations can be approximated by averaging over the corresponding terms using realizations of the unobserved variables (Liu, Wolfe, and Huang 2004). In the `strmcure` command, we adopt a numerical approximation method to implement Louis’s formula. To be specific, when the EM algorithm converges, we draw  $Q$  realizations for the two random vectors  $K = (k_1, \dots, k_n)^T$  and  $W = (\omega_1, \dots, \omega_n)^T$  according to their conditional probability density functions given in (4) and (5), given the parameter estimates  $\hat{\zeta}$  and the observed data  $\Omega$ . Denote the  $r$ th ( $r = 1, \dots, Q$ ) realization of the two random vectors as  $K^{(r)} = (k_1^{(r)}, \dots, k_n^{(r)})^T$  and  $W^{(r)} = (\omega_1^{(r)}, \dots, \omega_n^{(r)})^T$ , respectively. We evaluate the three terms  $\partial^2 \log L_c / \partial \zeta \partial \zeta^T$ ,  $\partial \log L_c / \partial \zeta$ , and  $\partial \log L_c / \partial \zeta^T$ , given the parameter estimate  $\hat{\zeta}$ , the observed data  $\Omega$ , and the  $r$ th realization  $K^{(r)}$  and  $W^{(r)}$ . The resultant three matrices are denoted by  $a_1^{(r)}$ ,  $a_2^{(r)}$ , and  $a_3^{(r)}$ , respectively. Then, the observed information matrix is estimated by

$$I(\hat{\zeta}) = \frac{-1}{Q} \sum_{r=1}^Q a_1^{(r)} - \frac{1}{Q} \sum_{r=1}^Q a_2^{(r)} + \left( \frac{1}{Q} \sum_{r=1}^Q a_3^{(r)} \right) \left( \frac{1}{Q} \sum_{r=1}^Q a_3^{(r)} \right)^T$$

The above evaluations were done using Stata’s matrix programming language, Mata, which provides computational benefits and supports many matrix operations.

We can then estimate the variance–covariance matrix of the parameter estimates by  $I^{-1}(\hat{\boldsymbol{\zeta}})$ . The dimension of  $I(\boldsymbol{\zeta})$  can be high because it depends on that of  $\mathbf{h}_0$ . Storing and inverting such a high-dimensional matrix is not always feasible. One feasible solution is to invert only the submatrix  $I_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\eta})$  to get the variance–covariance matrix estimate for  $\hat{\boldsymbol{\eta}}$ , where

$$I(\boldsymbol{\zeta}) = \begin{pmatrix} I_{\boldsymbol{\eta}\boldsymbol{\eta}}(\boldsymbol{\eta}) & I_{\boldsymbol{\eta}\mathbf{h}_0}(\boldsymbol{\eta}, \mathbf{h}_0) \\ I_{\boldsymbol{\eta}\mathbf{h}_0}(\boldsymbol{\eta}, \mathbf{h}_0) & I_{\mathbf{h}_0\mathbf{h}_0}(\mathbf{h}_0) \end{pmatrix}$$

This greatly simplifies the variance estimation procedure and has been discussed in Therneau and Grambsch (2000), Abrahantes and Burzykowski (2005), and Xu et al. (2012).

### 3.2 Tail-completion method

The baseline hazards ( $\mathbf{h}_0$ ) are estimated nonparametrically and are undefined beyond the observed largest uncensored event time. For a subject who experiences no event after the largest observed uncensored event time in the first event stratum (denoted by  $\tau_1^*$ ), there is no empirical basis for determining whether this means  $k_i = 0$  or small  $\omega_i$ . This near nonidentifiability problem may manifest itself as a flat or irregular likelihood surface (Sy and Taylor 2000). A simple solution is to impose a zero-tail constraint, which classifies all of those who are event free beyond  $\tau_1^*$  as being cured (for example,  $k_i = 0$ ).

Peng (2003) proposed using a parametric tail-completion method. This models the baseline hazard at the tail with a continuous function such that the baseline survival probability is a proper distribution for the uncured subjects. Two candidate continuous functions are the exponential and Weibull hazard functions. If the exponential function is used for the tail, after  $\tau_1^*$ , the baseline hazard remains at a constant level estimated by  $\hat{h}_{01}(\tau_1^*)$  using (9). If the Weibull function is used, the cumulative baseline hazard is modeled as  $(\alpha t)^\kappa$  for  $t > \tau_1^*$ . The two parameters  $\alpha$  and  $\kappa$  are estimated by maximum likelihood using all the observations on time to first event, subject to the constraint that  $(\hat{\alpha}\tau_1^*)^\kappa = \hat{H}_{01}(\tau_1^*)$ . All three methods performed well in various simulations (Xu et al. 2012).

Users can specify their choice on the tail-completion method by using the `tailcm()` option in the `strmcure` command. The default is `tailcm(zerotail)`.

## 4.1 Syntax

```
strmcure varlist [ if ] [ in ], shared(varname) [ strata(varname)
      xoffset(varname) zlist(varlist) zoffset(varname) znoconstant
      link(logistic|probit|cloglog) lastpool(#)
      tailcm(zerotail|exponential|weibull) phi0(#) iterate(#)
      tolerance(#) log baseh0(varname) saving seed(#) reps(#)
      repspart(#) dots level(#) obspart(#) rspart(#) ]
```

`strmcure` is for use with `st` data. Data must be `stset` before using this command. All the models include a gamma-distributed frailty term.

*varlist* may contain factor variables; see [U] 11.4.3 Factor variables.

`bootstrap`, `by`, `jackknife`, `statsby`, `stepwise`, and `xi` are allowed; see [U] 11.1.10 Prefix commands.

## 4.2 Options

`shared(varname)` specifies a variable in the dataset that identifies the subjects. It overrides the `id(varname)` specification in `stset` if it has been specified in `stset`. `shared()` is required.

`strata(varname)` stratifies the recurrent event times according to the specified event-order variable *varname*. Observations with the same value belong to the same stratum, and the baseline hazard function is unique to each stratum. Not specifying `strata()` means fitting a model without stratification when the model does not involve a cured fraction. This option must be specified for the frailty-mixture model.

`xoffset(varname)` specifies an offset variable for modeling the log(hazard). The regression coefficient is constrained to one.

`zlist(varlist)` specifies the list of variable names to be included in modeling the probability of being noncured in the frailty-mixture model.

`zoffset(varname)` specifies an offset variable for modeling the probability of being noncured in the frailty-mixture model. The regression coefficient is constrained to one.

`znoconstant` suppresses the constant term (intercept) when modeling the probability of being noncured in the frailty-mixture model.

`link(logistic|probit|cloglog)` specifies the link function for modeling the probability of being noncured in the frailty mixture model.

**lastpool(#)** specifies an integer  $j$  such that the  $j$ th and later events as specified by the *varname* in **strata(varname)** will be pooled to form one stratum.

**tailcm(zerotail | exponential | weibull)** specifies the tail-completion method for the frailty-mixture model. The default is **tailcm(zerotail)**.

**phi0(#)** specifies the starting value for the parameter (that is, the variance of the gamma-distributed frailty). The default is **phi0(2)**. If a negative value is specified, then the specification will be ignored and replaced with the default value of **phi0(2)**.

**iterate(#)** specifies the maximum number of iterations for the EM estimation. The default is **iterate(5000)**.

**tolerance(#)** specifies the minimum tolerance level for the regression coefficient estimation. The default is **tolerance(10<sup>-5</sup>)**. The **tolerance(#)** works together with **iterate(#)**: the EM estimation procedure is regarded as converged if either criterion is met.

**log** requests that the parameter estimates from each iteration of the EM estimation procedure be displayed.

**baseh0(varname)** requests that the estimated baseline hazard (using the Nelson–Aalen estimator) be saved. The variable *varname* is replaced if it already exists.

**saving** requests that the iterative parameter estimates from the EM estimation procedure be saved. The saved result is in the matrix **e(EMiter\_par)**. If the column or row of the matrix **e(EMite\_par)** exceeds the matsize limit, the user may not be able to read the matrix into Stata using the command **matlist e(EMiter\_par)**. In this case, the user can read the matrix into Mata or by converting the matrix to variables in a Stata dataset.

**seed(#)** sets the random-number seed when initiating the process of variance estimation using Louis’s formula. The default is **seed(0)**.

**reps(#)** specifies the number of realizations for the Monte Carlo simulation when estimating the variance using Louis’s formula. The specified number should be a multiple of 10. If not, it will be rounded up. The default is **reps(1000)**. Specifying **reps(0)** indicates no estimation of variance using Louis’s formula. Users can specify the **reps(0)** option if they use the **bootstrap** or **jackknife** method for variance estimation.

**repspart(#)** determines how many blocks to split the Monte Carlo replications into for the variance estimation using Louis’s formula. For each block,  $K = (\text{reps}/\text{repspart})$  realizations of the latent variables are generated. The default is **repspart(10)**. A larger  $K$  requires less computation time but a larger matrix size.

**dots** displays the replication dots in the estimation of variance using Louis’s formula. One dot is displayed for each successful completion with one block of  $K$  Monte Carlo realizations.

`level(#)` specifies the confidence level for estimating the confidence interval. The default is `level(95)`.

`obspart(#)` specifies the number of partitions for the total number of observations. This is used for partitioning the at-risk set in variance estimation using Louis's formula. The default is `obspart(N1)`, where `N1` is (total number of observations)/500, rounded up to the next integer. A smaller number requires less computation time but a larger matrix size.

`rspart(#)` specifies the number of partitions for the total number of events in variance estimation. The default is `rspart(N2)`, where `N2` is (total number of events)/500, rounded up to the next integer. A smaller number requires less computation time but a larger matrix size.

## 4.3 Stored results

`strmcure` stores the following in `e()`:

### Scalars

<code>e(N)</code>	number of observations
<code>e(N_sub)</code>	number of subjects
<code>e(N_fail)</code>	number of failures
<code>e(N_g)</code>	number of groups
<code>e(g_min)</code>	smallest group size
<code>e(g_max)</code>	largest group size
<code>e(g_avg)</code>	average group size
<code>e(total_time)</code>	total analysis time
<code>e(CurePresent)</code>	whether a cured fraction id specified (0 no and 1 yes)
<code>e(phi_initial)</code>	initial value for the parameter phi (that is, the variance of the gamma-distributed frailty)
<code>e(tol_v_cvg)</code>	the square root of the distance between the two parameter estimate vectors at the last iteration and its preceding iteration of the EM estimation procedure
<code>e(iter_cvg)</code>	the total number of iterations at the convergence of the EM estimation procedure
<code>e(Tail)</code>	code number of the tail-completion method for the frailty-mixture model

### Macros

<code>e(cmd)</code>	<code>strmcure</code>
<code>e(cmdline)</code>	command as typed
<code>e(depvar)</code>	name of dependent variable
<code>e(title)</code>	frailty model or frailty-mixture model
<code>e(offsetX)</code>	name of the offset variable for modeling the log(hazard)
<code>e(offsetZ)</code>	name of the offset variable for modeling the probability of being noncured
<code>e(Tailname)</code>	tail-completion method for the frailty-mixture model
<code>e(link)</code>	name of the link function for the frailty-mixture model
<code>e(ties)</code>	method used for handling ties (Breslow method is always used)
<code>e(shared)</code>	frailty grouping variable
<code>e(t0)</code>	<code>_t0</code>
<code>e(properties)</code>	<code>b V</code>
<code>e(strata)</code>	strata variable

Matrices	
<b>e(b)</b>	coefficient vector
<b>e(V)</b>	variance–covariance matrix of the estimators
<b>e(EMiter_par)</b>	matrix storing the history of the parameter estimates during the iterative EM estimation procedure, if <b>saving</b> option was specified in the <b>strmcure</b> command
Functions	
<b>e(sample)</b>	marks estimation sample

## 5 Example

Here we use the rhDNase trial of respiratory exacerbations described in Cook and Lawless (2007) (<http://www.math.uwaterloo.ca/~rjcook/book/example2/rhDNase.dat>). This study was a randomized trial on the effect of rhDNase versus a placebo. rhDNase was an enzyme that was supposed to digest the harmful extracellular DNA that accumulated in the lungs of patients with cystic fibrosis. It was hypothesized that the rhDNase group would have a lower incidence of respiratory exacerbations. The 645 patients had 0 to 5 exacerbations during the average 166 days of follow-up.

Some suspected errors are in the data. We start by modifying these data.

```
. use data_rhdnase
. recode etype 1=2 2=1 if id==951319 | id==985308 | id==985316 | id==986310
(etype: 14 changes made)
```

We first demonstrate the command using the counting-process time scale.

```
. stset time2 if etype==1, fail(status) id(id) time0(time1) exit(time .)
(output omitted)
. sort id _t0
. list id enum etype _st _d _t _t0
```

	id	enum	etype	_st	_d	_t	_t0
1.	493301	1	1	1	0	168	0
2.	493303	1	1	1	0	169	0
3.	493305	1	1	1	1	65	0
4.	493305	3	1	1	0	168	75
5.	493305	2	2	0	.	.	.

In this dataset, **etype=2** represents the acute treatment duration initiated by exacerbation, during which the patient was not at risk of another exacerbation. Therefore, the gap (and missing value) in **id** 493305 occurs between day 65 and 75. We drop records with **etype=2** from now on.

### No cured fraction, no stratification

This model is conceptually identical to the shared frailty model of **stcox**, which uses profile likelihood. However, **strmcure** is much faster. We run both commands in Stata/SE 12.1 for Windows (64-bit) in a PC with Duo CPU each at 2.50 GHz. We use the

command `set rmsg on` to show the number of seconds needed to run the models. The log hazard-ratio (HR) and the standard error from the two commands are very similar, but `strmcure` takes only a fraction of the time `stcox` does (that is, 9.91 seconds versus 223.30 seconds). This is because in `stcox`, the subject-specific frailty terms are treated as regression coefficients of indicators specific to each subject; thus the dimension of the parameter space to be estimated depends on the sample size. As such, the maximum likelihood estimation involves the inversion of a high-dimensional Hessian matrix for the parameters. This could slow down the estimation process. There is no such problem when using the `strmcure` command, because the parameter estimation was done by using an EM algorithm, which treats the random effects as missing variables. Because there is no stratification for event order, the log HR  $-0.269$  is an estimate of the total effect. `rhDNase` is associated with a reduction of  $\{1 - \exp(-0.269)\} = 24\%$  in event rate.

```
. strmcure trt, shared(id) dots
      (output omitted)
Frailty model -- No stratification
      Breslow method for ties          Number of obs      =          965
      Gamma shared frailty            Number of groups    =          645
Group variable: id
No. of subjects      =          645      Obs per group: min =          1
No. of failures      =          364      avg =    1.496124
Total analysis time  =         101156    max =          5
```

_t		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Beta							
	trt	-.2686094	.1400622	-1.92	0.055	-.5431264	.0059075
frailty							
	_cons	1.281714	.2445497	5.24	0.000	.8024052	1.761022

```
. query memory
      (output omitted)
. set matsize 700
. stcox trt, shared(id) nohr forceshared
      (output omitted)
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
trt	-.2686098	.1396921	-1.92	0.054	-.5424014	.0051817
theta	1.281589	.2315933				

```
Likelihood-ratio test of theta=0: chibar2(01) =    71.53 Prob>=chibar2 = 0.000
Note: Standard errors of regression parameters are conditional on theta.
. set matsize 400
```



## No cured fraction, stratification

We then rerun the model but with stratification for event order. Note that currently, `stcox` does not allow the use of both the `shared()` and `strata()` options, so this model cannot be fit in `stcox`. We pool the fourth and fifth events into the same stratum because of sparse data. The log HR is  $-0.412$ , or  $[1 - \exp(-0.412)] = 34\%$  reduction in event rate. This estimates the primary effect. Comparing the two log HRs suggests negative event dependence.

```
. stgen event_order=nfailures()
(361 missing values generated)
. replace event_order=event_order+1
(965 real changes made)
. strmcure trt, shared(id) strata(event_order) lastpool(4) dots
(output omitted)
```

		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Beta	_t						
	trt	-.4123722	.2397462	-1.72	0.085	-.8822662	.0575217
frailty	_cons	5.814084	.5602115	10.38	0.000	4.716089	6.912078

## Cured fraction, stratification

Next, we use gap times to illustrate, as follows:

```
. generate time=time2-time1
. stset time if etype==1, fail(status) exit(time .)
(output omitted)
. sort id _t
. list id enum etype _st _d _t _t0
```

	id	enum	etype	_st	_d	_t	_t0
1.	493301	1	1	1	0	168	0
2.	493303	1	1	1	0	169	0
3.	493305	1	1	1	1	65	0
4.	493305	3	1	1	0	93	0
5.	493309	1	1	1	0	168	0

We plot the Kaplan–Meier survival estimate of time to the first respiratory exacerbation (figure 1).

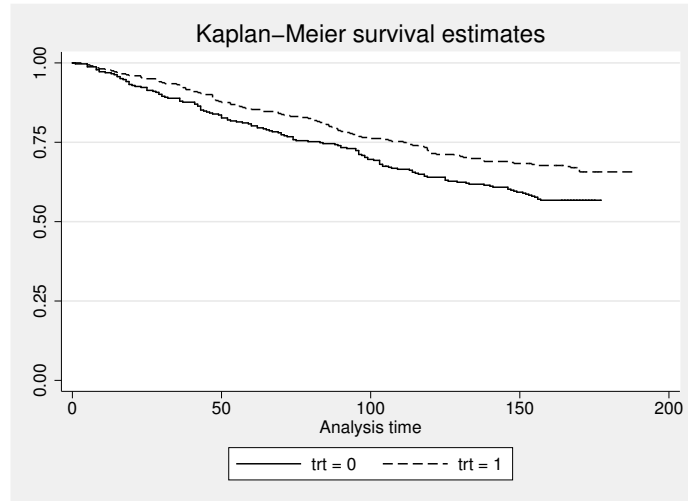


Figure 1. Kaplan–Meier survival estimate of time to the first respiratory exacerbation

```
. sts graph if event_order==1, by(trt)
```

We see in figure 1 that the survival curves in the two study groups appear to plateau sometime before the study closure. This empirical observation from the data suggested that a cured fraction might exist in the study population that was not at risk in the study event. We then fit a model with stratification and a cured fraction. We use the logit link function and zero-tail constraint for tail completion.

```
. stmcmure trt, shared(id) strata(event_order) lastpool(4) zlist(trt)
> link(logistic) tailcm(ze) dots
(output omitted)
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Theta						
trt	-.4455727	.1722883	-2.59	0.010	-.7832516	-.1078939
_cons	-.1685116	.1242365	-1.36	0.175	-.4120107	.0749875
Beta						
trt	.0576491	.2064493	0.28	0.780	-.3469842	.4622823
frailty						
_cons	.8088558	.1596303	5.07	0.000	.4959861	1.121725

The log odds-ratio for being noncured is  $-0.446$ . The log HR is close to zero. The effect of rhDNase on exacerbations is mainly through making more patients exacerbation free. To use the terminology in infection disease prevention, we see that rhDNase appears to work in an all-or-none model. Finding predictors of who may benefit from the intervention is particularly important in this situation.

Furthermore, we investigate the effect of rhDNase after adjustment for forced expiratory volume at randomization (FEV). FEV measures a person's lung function.

```
. strmcure trt fev, shared(id) strata(event_order) lastpool(4) zlist( trt fev)
> link(logistic) tailcm(ze) dots
(output omitted)
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
Theta						
trt	-.488413	.1812871	-2.69	0.007	-.8437292	-.1330968
fev	-.0272301	.0038077	-7.15	0.000	-.0346931	-.0197671
_cons	1.465944	.2655137	5.52	0.000	.9455466	1.986341
Beta						
trt	.0513148	.1695338	0.30	0.762	-.2809653	.3835949
fev	-.0045154	.0041625	-1.08	0.278	-.0126738	.003643
frailty						
_cons	.8271226	.2051831	4.03	0.000	.4249712	1.229274

The estimation results for rhDNase are similar to those without adjustment for FEV. This is expected in a randomized trial because the effect of a randomized intervention should not be confounded by baseline covariates. After adjusting for the intervention effect, we see that a higher FEV is associated with a lower probability of experiencing any respiratory exacerbation, with log odds-ratio  $-0.027$  per FEV unit.

We do not have a biological explanation for the possibility of a cured fraction in the study population in this example. The analyses were mainly used to illustrate the use of the macro.

## 6 Conclusion

We are interested in the estimation of the dual aspects of the intervention effect in reducing the subjects' probability of being susceptible and reducing the event hazard (but not to zero) within the susceptible subjects. If used appropriately, models that account for the coexistence of a cured fraction and unobserved heterogeneity among those noncured can be informative. An important assumption underlying such models is the existence of a cured fraction in the population. The existence of a cured fraction cannot be tested by statistical means alone. Rather, both empirical evidence (for example, the survival curve or cumulative hazard curve reaching a plateau) and biological justification or other external evidence must be in place before one can confidently apply such models and draw a conclusion from the study results. When the evidence and biological

justification for the existence is limited, one may consider using a simpler model that accounts for only the unobserved heterogeneity.

Previous research has accounted for a cured fraction or unobserved heterogeneity among the subjects in the study population, but not both: for example, the “all-or-none” model versus the “leaky” model in vaccine research (Halloran, Longini, and Struchiner 1996) and the “split population” survival time model (Schmidt and Witte 1989) in econometric literature. The all-or-none model and split-population model assume that the study population consists of a cured proportion of subjects who will never fail, whereas the leaky model assumes all study subjects will eventually fail with varying risks. In reality, it may be the combination of both; the study population may consist of subjects who will never fail, who are at low risk of failure, and who are at relatively higher risk of failure. This requires the underlying subject-specific frailty to have a point mass at zero and a continuous part that takes positive values. In this situation, the proposed frailty-mixture model is potentially useful because it provides a hybrid of these models. In the frailty-mixture model, the point mass at zero in the frailty (that is, the cured fraction) is modeled using a binary regression function with three common choices for the link functions: the logistic, probit, or complementary log-log link. The continuous positive part of frailty for the noncured is modeled by a Gamma distribution. The choice of Gamma distribution for modeling the frailty among the noncured was mainly because of its mathematical convenience. Other distributional forms for the positive part of frailty could be used, such as inverse Gaussian and positive stable distributions (Hougaard 1986). The robustness of the proposed model estimation to a misspecified frailty distribution has been demonstrated by our previous work (Xu et al. 2012).

The `strmcure` command provides a useful alternative to the existing `stcox` command for the analysis of recurrent event data without a cured fraction. Furthermore, it facilitates a deeper understanding of intervention effects by allowing a cured fraction. The `strmcure` command implements the estimation of frailty-mixture models. In essence, the frailty-mixture models extend the Cox-type proportional hazard models by incorporating a cured fraction as well as a subject-specific frailty multiplicative to the hazard for the noncured. In the hazard function for a noncured subject as specified in (2), the baseline hazard functions are left unspecified, whereas the effects of covariates (modeled via an exponential function) and subject-specific frailty are both multiplicative to the baseline hazard. The advantage of this semiparametric specification is that one does not have to specify the functional form for the baseline hazard. Alternatively, one may be concerned about the assumption of constant hazard ratio (or the proportionality of hazard) implied by (2) being too restrictive for real-life applications. As noted by Allison (2010, 172), the proportionality assumption can hardly be satisfied exactly, just as any other statistical assumptions. A working solution for applying the frailty-mixture models with nonproportionality is to incorporate an interaction term between the covariate concerned and (some functional form of) time in the model to capture the time-varying covariate effect. However, we agree with Allison (2010, 155) that it is usually all right to ignore violation of the proportional hazard assumption. In that case, the hazard ratio represents the average effect of the covariate over study time in the data, which is useful.

The `strmcure` command handles ties using the Breslow approximation method. If the number of ties relative to the at-risk set is large, one may consider breaking the ties by using the jittering method (Tai et al. 2002) in the following three steps: 1) add or subtract a small number to the ties; 2) apply the `strmcure` command to the dataset; and 3) repeat steps 1 and 2 several times, and then combine the multiple parameter estimates using Rubin’s rule (Rubin 1987; Tai et al. 2002). Furthermore, when the data encounter a substantial number of ties or are broadly grouped, it is more natural to use discrete-time analysis models. Therefore, among others, one valuable consideration for future work is to extend the proposed model to accommodate discrete-time data to account for the variations in the individual risks as well as for a cured fraction within the population.

## 7 Acknowledgment

This work was supported by the Singapore Ministry of Health’s National Medical Research Council under its Clinician Scientist Award.

## 8 References

- Aalen, O. O. 1988. Heterogeneity in survival data analysis. *Statistics in Medicine* 7: 1121–1137.
- Abrahantes, J. C., and T. Burzykowski. 2005. A version of the EM algorithm for proportional hazard model with random effects. *Biometrical Journal* 47: 847–862.
- Allison, P. D. 2010. *Survival Analysis Using SAS: A Practical Guide*. 2nd ed. Cary, NC: SAS Institute.
- Box-Steffensmeier, J. M., and S. De Boef. 2006. Repeated events survival models: The conditional frailty model. *Statistics in Medicine* 25: 3518–3533.
- Cheung, Y. B., Y. Xu, S. H. Tan, F. Cutts, and P. Milligan. 2010. Estimation of intervention effects using first or multiple episodes in clinical trials: The Andersen-Gill model re-examined. *Statistics in Medicine* 29: 328–336.
- Cook, R. J., and J. F. Lawless. 2007. *The Statistical Analysis of Recurrent Events*. New York: Springer.
- Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 39: 1–38.
- Greenland, S. 1996. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology* 7: 498–501.
- Halloran, M. E., I. M. Longini, Jr., and C. J. Struchiner. 1996. Estimability and interpretation of vaccine efficacy using frailty mixing models. *American Journal of Epidemiology* 144: 83–97.

- Hougaard, P. 1986. A class of multivariate failure time distributions. *Biometrika* 73: 671–678.
- Keiding, N., P. K. Anderson, and J. P. Klein. 1997. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Statistics in Medicine* 16: 215–224.
- Kelly, P. J., and L. L. Lim. 2000. Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine* 19: 13–33.
- Liu, L., R. A. Wolfe, and X. L. Huang. 2004. Shared frailty models for recurrent events and a terminal event. *Biometrics* 60: 747–756.
- Louis, T. A. 1982. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B* 44: 226–233.
- Metcalfe, C., and S. G. Thompson. 2006. The importance of varying the event generation process in simulation studies of statistical methods for recurrent events. *Statistics in Medicine* 25: 165–179.
- Moorthy, V. S., Z. R. Reed, and P. G. Smith. 2009. MALVAC 2008: Measures of efficacy of malaria vaccines in phase 2b and phase 3 trials—scientific, regulatory and public health perspectives. *Vaccine* 27: 624–628.
- O’Quigley, J., and J. Stare. 2002. Proportional hazards models with frailties and random effects. *Statistics in Medicine* 21: 3219–3233.
- Peng, Y. W. 2003. Estimating baseline distribution in proportional hazards cure models. *Computational Statistics and Data Analysis* 42: 187–201.
- Pickles, A., and R. Crouchley. 1994. Generalizations and applications of frailty models for survival and event data. *Statistical Methods in Medical Research* 3: 263–278.
- . 1995. A comparison of frailty models for multivariate survival data. *Statistics in Medicine* 14: 1447–1461.
- Prentice, R. L., B. J. Williams, and A. V. Peterson. 1981. On the regression analysis of multivariate failure time data. *Biometrika* 68: 373–379.
- Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Schmidt, P., and D. Witte. 1989. Predicting criminal recidivism using split population survival time models. *Journal of Econometrics* 40: 141–159.
- Simpson, S. E. 2013. A positive event dependence model for self-controlled case series with applications in postmarketing surveillance. *Biometrics* 69: 128–136.
- Smith, P. G., L. C. Rodrigues, and P. E. Fine. 1984. Assessment of the protective efficacy of vaccines against common diseases using case-control and cohort studies. *International Journal of Epidemiology* 13: 87–93.

- Sy, J. P., and J. M. G. Taylor. 2000. Estimation in a Cox proportional hazards cure model. *Biometrics* 56: 227–236.
- Tai, B. C., I. R. White, V. Gebski, and D. Machin. 2002. On the issue of ‘multiple’ first failures in competing risks analysis. *Statistics in Medicine* 21: 2243–2255.
- Therneau, T. M., and P. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer.
- Ullah, S., T. J. Gabbett, and C. F. Finch. 2014. Statistical modelling for recurrent events: An application to sports injuries. *British Journal of Sports Medicine* 48: 1287–1293.
- Xu, Y., Y. B. Cheung, K. F. Lam, and P. Milligan. 2012. Estimation of summary protective efficacy using a frailty mixture model for recurrent event time data. *Statistics in Medicine* 31: 4023–4039.

### **About the authors**

Ying Xu is an assistant professor in the Center for Quantitative Medicine at Duke–NUS Graduate Medical School, Singapore. Her research interests include statistical methodology related to infectious disease and human growth as well as design and analysis of clinical trials.

Yin Bun Cheung is a medical statistician and pediatric epidemiologist. He is a professor in the Center for Quantitative Medicine at Duke–NUS Graduate Medical School, Singapore, and an adjunct professor in the Department of International Health at the University of Tampere, Finland. His current research areas include statistical methods for analysis of censored data and excess zeros and a nonsusceptible fraction and the impact and interplay of infection and undernutrition on maternal and child health in developing countries.