The Stata Journal (2022)
22, Number 3, pp. 679–701 DOI: 10.1177/1536867X221124552

Flexible parametric survival analysis with multiple timescales: Estimation and implementation using stmt

Hannah Bower
Clinical Epidemiology Division
Department of Medicine Solna, Karolinska Institutet
Stockholm, Sweden
hannah.bower@ki.se

Therese M.-L. Andersson
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
therese.m-l.andersson@ki.se

Michael J. Crowther
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
michael.crowther@ki.se

Paul C. Lambert
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
and
Biostatistics Research Group
Department of Health Sciences, University of Leicester
Leicester, U.K.
paul.lambert@leicester.ac.uk

Abstract. In this article, we describe methodology that allows for multiple timescales using flexible parametric survival models without the need for time splitting. When one fits flexible parametric survival models on the log-hazard scale, numerical integration is required in the log likelihood to fit the model. The use of numerical integration allows incorporation of arbitrary functions of time into the model and hence lends itself to the inclusion of multiple timescales in an appealing way. We describe and exemplify these methods and show how to use the command `stmt`, which implements these methods, alongside its postestimation commands.

Keywords: st0688, stmt, stmt postestimation, flexible parametric survival model, multiple timescales

1 Introduction

Defining the timescales of interest is essential when performing any time-to-event analysis. Most commonly, only one timescale of interest is modeled using standard survival analysis methods. There may be, however, certain situations where modeling multiple timescales is preferred or necessary to obtain useful results and answer research questions of interest (Iacobelli and Carstensen 2013). Consider the example of time to death after a cancer diagnosis. Here two timescales may be of importance: time since diagnosis and attained age. Time from cancer diagnosis to death is a highly important timescale and is commonly used in studies of this type (although this does vary according to the cancer diagnosis under study). However, it is also known that age is an important factor to consider for the risk of death. In studies of breast cancer incidence, attained age and time since childbirth are two important timescales (Albrektsen et al. 2006). In studies of infection risk after admittance to intensive care unit, both time in intensive care and calendar time are important timescales to consider (Wolkewitz et al. 2016). In the field of engineering, with regard to studying engine failure, attained age of the engine and the usage (time) of the engine are valuable timescales (Duchesne and Lawless 2000).

When more than one timescale is thought to be important in an exposure-outcome association, it can be common for analysts to simplify their approach by choosing the most important timescale to model and to select and model some time-fixed version of a second timescale rather than modeling, for example, in age-period-cohort models (Carstensen 2007). If modeling the hazard as a function of the first timescale is only of interest and the second timescale is sufficiently captured using the time-fixed version, then this is a reasonable approach. This is often the case when considering time to death after a cancer diagnosis; the hazard is often modeled as a function of time since diagnosis and age at diagnosis. However, if modeling the hazard as a function of both timescales is of interest, then it may be important to capture the effect of multiple timescales simultaneously. For example, it may be of interest to model the mortality rate as a function of both time since diagnosis and attained age in cancer patients. It has also been shown that effect estimates may be biased in some situations where the underlying hazard is a function of multiple timescales (Batyrbekova et al. Forthcoming).

In situations where one would like to model the hazard as a function of multiple timescales, it is common to split the second timescale and either 1) model the hazard as a step function of the second timescale or 2) smooth this function by including, for example, a spline (Carstensen 2006; Royston and Lambert 2011). Note that splitting across timescales can be computationally intensive depending on the size of the study population and the number of splits across the timescale one wishes to make. Modeling time as a continuous function, for example, by using splines, has been shown to be able to equally or more accurately capture an array of hazard functions that other methods based on parametric assumptions may struggle with (Royston and Lambert 2011).

Because time increases at the same rate independently of the timescale, it is possible to present each timescale as a function of another (Efron 2002; Danardono 2005; Iacobelli and Carstensen 2013). The difference between two timescales is simply a difference between the origins of the timescales. For example, the difference between the timescales time since diagnosis and attained age is simply age at diagnosis. We use this approach to model multiple timescales in a flexible parametric survival-model framework. Our previous article describes the command `strcs`, which fits flexible parametric survival models on the log-hazard scale using numerical integration to obtain the log likelihood (Bower, Crowther, and Lambert 2016). The use of numerical integration allows incorporation of arbitrary functions of time into the model and hence lends itself to the inclusion of multiple timescales in an appealing way. We now extend this approach to model multiple timescales and introduce the command `stmt`, which models multiple timescales using flexible parametric survival models on the log-hazard scale. In section 2, we describe flexible parametric survival models in general and then, in section 3, describe how multiple timescales can be implemented in these models. In sections 4 and 5, we describe the command `stmt` and its postestimation commands. Finally, in section 6 we present an illustrative example of modeling multiple timescales using the `stmt` command.

2 Flexible parametric survival models

A flexible parametric survival model on the log-hazard scale that assumes proportional hazards can be written as

$$\ln\{h(t|\boldsymbol{x})\} = s\{f(t)|\gamma, \boldsymbol{k}\} + \boldsymbol{x}\boldsymbol{\beta} \tag{1}$$

where $s\{f(t)|\gamma, \boldsymbol{k}\}$ represents the restricted cubic spline function that forms the baseline log-hazard function for time t with knots $\boldsymbol{k}$ (further details on restricted cubic splines can be found in Bower, Crowther, and Lambert [2016]) and $\boldsymbol{\beta}$ represents the log-hazard ratio estimates for variables contained in $\boldsymbol{x}$. Estimates from flexible parametric survival models have been shown to be approximately the same as estimates from a similar Cox model (Royston and Lambert 2011; Rutherford, Crowther, and Lambert 2015). It is common for the spline function to have $f(t) = t$ or $f(t) = \ln(t)$.

The flexible parametric survival model is easily extended to include time-dependent effects and hence relax the proportional-hazards assumption by introducing an interaction between $\boldsymbol{x}_y$ and a spline term:

$$\ln\{h(t|\boldsymbol{x})\} = s\{f(t)|\gamma, \boldsymbol{k}\} + \boldsymbol{x}\boldsymbol{\beta} + \sum_{y=1}^{Y} s\{f(t)|\gamma_y, \boldsymbol{k}_y\} \times \boldsymbol{x}_y \tag{2}$$

Y is the number of time-dependent effects, and $s\{f(t)|\gamma_y, \boldsymbol{k}_y\}$ is the spline function for the yth time-dependent effect.

2.1 Maximum likelihood estimation

Flexible parametric survival models are fit using maximum likelihood estimation; the `ml` command is used in Stata. Consider a sample of n individuals who are followed over time t_i; d_i represents the event indicator for the ith individual. Then the log-likelihood contribution for the ith individual is

$$\log l_i = d_i \log\{h(t_i)\} + \log\{S(t_i)\} \tag{3}$$

where $h(t_i)$ is the hazard function and $S(t_i)$ is the survival function evaluated at the time of event or censoring t_i. Using the relationship between the survival function and the cumulative hazard function,

$$S(t_i) = \exp\{-H(t_i)\}$$

we can rewrite the log likelihood in (3) as

$$\log l_i = d_i \log\{h(t_i)\} - H(t_i) \tag{4}$$

Under delayed entry, where individual i becomes at risk at t_{0i}, (4) becomes

$$\log l_i = d_i \log\{h(t_i)\} - H(t_i) + H(t_{0i}) \tag{5}$$

Thus, when fitting a flexible parametric survival model, we need the hazard $h(t_i)$ and the cumulative hazard $H(t_i)$ to evaluate the likelihood. Under delayed entry, $H(t_{0i})$ is also required.

Consider the flexible parametric survival model in (2). Theoretically, components of the log likelihood presented in (4) can be calculated because the log-hazard function is being modeled and because of the following relationship:

$$H(t_i) = \int_{t_{0i}}^{t_i} h(u_i)\ du$$

However, because restricted cubic splines are used to estimate the log baseline hazard function, the hazard function cannot be integrated analytically. Thus, numerical integration techniques must be used to calculate the cumulative hazard function and evaluate the likelihood. In this application, Gaussian quadrature is used to numerically integrate the hazard function; this converts an integral into a weighted sum over a set of predefined points called nodes. This is further described in section 3.1.

3 Modeling multiple timescales using flexible parametric survival models on the log-hazard scale

Because time progresses at the same rate, multiple timescales can be considered as a function of one another. Thus, the difference between timescales is really a question of the difference between the origins of the timescales. If we consider the situation where there are two timescales of interest, where t_1 and t_2 represent the first and second timescales, respectively, there exists some constant c that

$$t_2 = t_1 + c$$

For example, if we consider t_1 to be time since diagnosis and t_2 to be attained age, then our constant c would be age at diagnosis, the difference between the two timescales. This relationship extends to multiple timescales, where all timescales can be written as a function of the first timescale and some constant. When modeling multiple timescales using flexible parametric survival models, we apply the same theory. Modeling multiple timescales using other types of survival models is possible, but the flexible parametric survival-model approach described here avoids the need for time splitting, which can be time consuming, and additionally allows us to simultaneously model the hazard as a function of two smooth timescales. Consider again the situation where one is interested in modeling the hazard as a function of two timescales simultaneously; this is in fact a special case of the bivariate hazard model whereby

$$\begin{aligned} h(t|x) &= h_0(t_1, t_2)\text{exp}(\boldsymbol{x\beta}) \\ &= h_{01}(t_1)h_{02}(t_2)\text{exp}(\boldsymbol{x\beta}) \\ &= h_{01}(t_1)h_{02}(t_1 + c)\text{exp}(\boldsymbol{x\beta}) \end{aligned} \tag{6}$$

Using splines to model the baseline hazard functions in (6) similarly to that shown in (1), we get the flexible parametric survival model with two timescales:

$$\begin{aligned} \ln\{h(t_1|\boldsymbol{x})\} &= s\{f(t_1)|\boldsymbol{\gamma}_1, \boldsymbol{k}_1\} + s\{f(t_1 + c)|\boldsymbol{\gamma}_2, \boldsymbol{k}_2\} + \boldsymbol{x\beta} \\ &= s\{f(t_1)|\boldsymbol{\gamma}_1, \boldsymbol{k}_1\} + s\{f(t_2)|\boldsymbol{\gamma}_2, \boldsymbol{k}_2\} + \boldsymbol{x\beta} \end{aligned} \tag{7}$$

Now there are two restricted cubic spline functions; $s\{f(t_1)|\boldsymbol{\gamma}_1, \boldsymbol{k}_1\}$ is a function of the first timescale with knots $\boldsymbol{k}_1$, whereas $s\{f(t_2)|\boldsymbol{\gamma}_2, \boldsymbol{k}_2\}$ is a function of the second timescale with knots $\boldsymbol{k}_2$. This concept extends to modeling several timescales.

3.1 Log likelihood and numerical integration

As previously described, maximum likelihood is used to estimate parameters in the flexible parametric survival model. Because the log-hazard function is modeled, the cumulative hazard should be calculated as shown in (4) and (5). Consider the model shown in (7), which models two timescales simultaneously. The cumulative hazard evaluated between t_0 and t can be written as

$$\begin{aligned} H(t) &= \int_{t_0}^{t} h(u)\ \text{du} \\ &= \int_{t_0}^{t} \exp\left[s\left\{f(u_1)|\boldsymbol{\gamma}_1, \boldsymbol{k}_1\right\} + s\left\{f(u_2)|\boldsymbol{\gamma}_2, \boldsymbol{k}_2\right\} + \boldsymbol{x\beta}\right]\ du \end{aligned} \tag{8}$$

Unfortunately, the restricted cubic splines $s\{f(t_1)|\boldsymbol{\gamma}_1, \boldsymbol{k}_1\}$ and $s\{f(t_2)|\boldsymbol{\gamma}_2, \boldsymbol{k}_2\}$ cannot be integrated analytically. Instead, Gauss–Legendre quadrature is implemented (Stoer and Bulirsch 2002) as follows, which converts an integral into a weighted summation over a set of predefined points known as nodes, m, across a function $g(u)$:

$$\int g(u)du \approx \sum_{j=1}^{m} w_j g(u_j)$$

If we are interested in estimating the integral between a and b, then this formula becomes

$$\int_a^b g(u)du \approx \frac{b-a}{2} \sum_{j=1}^{m} w_j g\left(\frac{b-a}{2}u_j + \frac{a+b}{2}\right) \tag{9}$$

Now considering that we are interested in estimating the integral of the hazard function between t_0 and t as displayed in (8), we can use (9) to get

$$\begin{aligned}
&\int_{t_0}^{t} \exp\left[s\left\{f(u_1)|\boldsymbol{\gamma}_1, \boldsymbol{k}_1\right\} + s(f(u_2)|\boldsymbol{\gamma}_2, \boldsymbol{k}_2) + \boldsymbol{x\beta}\right]\; du \\
=&\int_{t_0}^{t} \exp\left[s\left\{f(u_1)|\boldsymbol{\gamma}_1, \boldsymbol{k}_1\right\} + s\left\{f(u_1+c)|\boldsymbol{\gamma}_2, \boldsymbol{k}_2\right\} + \boldsymbol{x\beta}\right]\; du \\
\approx&\frac{t-t_0}{2}\sum_{j=1}^{m} \exp\left[s\left\{f\left(\frac{t-t_0}{2}u_{1j} + \frac{t_0+t}{2}\right)|\boldsymbol{\gamma_1}, \boldsymbol{k}_1\right\}\right. \\
&\qquad \left. +s\left\{f\left(\frac{t-t_0}{2}u_{1j} + \frac{t_0+t}{2} + c\right)|\boldsymbol{\gamma_2}, \boldsymbol{k}_2\right\} + \boldsymbol{x\beta}\right]
\end{aligned}$$

The summation is calculated over timescale 1, t_1, at m nodes. Note that the weights in (9) are set to 1 here. Because every timescale can be written as a function of the first timescale, the hazard function can be numerically integrated over the first timescale only, and the likelihood function can be calculated according to t_1. The number of quadrature nodes is selected by the user; the number of nodes required is dependent on the complexity of the hazard function, but it has been shown that approximately 30 nodes should be sufficient in the majority of situations (Crowther and Lambert 2014).

4 The stmt command

The `stmt` command fits flexible parametric survival models on the log-hazard scale while allowing multiple timescales (up to three) to be fit simultaneously. Restricted cubic splines smooth the log hazard with user-specified degrees of freedom. The first timescale is specified using the `stset` command as in other standard survival analyses. Additional timescales are included using the options described below. Covariates can be included within the model, and interactions between covariates and the timescale can be specified.

Numerical integration of the hazard function is undertaken via integration with Gauss–Legendre quadrature. Both the `rcsgen` (Lambert 2008) and `stpm2` (Lambert 2010) commands are called in `stmt` to create splines and obtain initial values, respectively; the user must install these prior to using the `stmt` command. The log likelihood is maximized using the Newton–Raphson algorithm via the `ml` command in Stata, using analytic derivatives for the score and Hessian to increase speed and accuracy.

4.1 Syntax

`stmt` *varlist* [*if*] [*in*] `, time1(`*suboptions*`)` [`time2(`*suboptions*`) time3(`*suboptions*`)`
`timeint(`*int_list*`) timeintknots(`*int_list*`) timeintbknots(`*int_list*`)`
`noconstant nodes(`*#*`) noorthog nohr verbose from(`*matrix*`) inith(`*varname*`)`
maximize_options]

4.2 Options

`time1(`*suboptions*`)` contains suboptions for timescale 1 (see below for a list of suboptions). The first timescale is always specified using the `stset` command. `time1()` is required.

`time2(`*suboptions*`)` contains suboptions for timescale 2 (see below for a list of suboptions). The second timescale is a function of the first timescale; the difference between the second timescale and the first timescale is specified in the `start()` suboption.

`time3(`*suboptions*`)` contains suboptions for timescale 3 (see below for a list of suboptions). The third timescale is a function of the first timescale; the difference between the third timescale and the first timescale is specified in the `start()` suboption.

`timeint(`*int_list*`)` specifies two-way timescale interactions. The syntax looks as follows: `timeint(t1:t2 2:4)`, where an interaction between timescale 1 and timescale 2 will be created. Restricted cubic splines with 2 degrees of freedom for timescale 1 will be interacted with restricted cubic splines with 4 degrees of freedom for timescale 2. The space separates the specified timescales and their degrees of freedom. Additional timescale interactions can be added using `|` as follows: `timeint(t1:t2 2:4 | t1:t3 2:2)`.

`timeintknots(`*int_list*`)` specifies the internal knots for two-way timescale interactions. The syntax looks as follows: `timeintknots(2 5 : 50 60 70)`, where an interaction between the timescales specified in `timeint()` will be created. This will create restricted cubic splines for the first timescale specified in the `timeint()` option with internal knots at 2 and 5 on this timescale (3 degrees of freedom) and restricted cubic splines for the second timescale specified in the `timeint()` option at 50, 60, and 70 on this timescale (4 degrees of freedom) and then interact the spline terms together. The spaces separate the knot locations, whereas the colon separates the timescales included in `timeint()`; that is, the internal knots for the first timescale in `timeint()` are specified prior to the colon, and the internal knots for the second timescale in `timeint()` come after the colon. Internal knots for additional timescale interactions can be added using `|` in a similar way to the `timeint()` option, for example, `timeintknots(2 5 : 50 60 70 | 2 3 5 : 2)`. Note that knots should be specified on the untransformed timescale.

`timeintbknots(`*int_list*`)` specifies the boundary knots for two-way timescale interactions. The syntax looks like `timeintbknots(timeint(0 7 : 25 95))`, where an interaction between the timescales specified in `timeint()` will be created. Restricted cubic splines with 2 degrees of freedom for the first timescale specified in `timeint()` will be interacted with restricted cubic splines with 4 degrees of freedom for the second timescale specified in `timeint()`. The spaces separate the boundary knot locations, whereas the colon separates the timescales included in `timeint()`; that is, the boundary knots for the first timescale in `timeint()` are specified prior to the colon, and the boundary knots for the second timescale in `timeint()` come after the colon. Boundary knots for additional timescale interactions can be added using `|` in a similar way to the `timeint()` option, for example, `timeintbknots(0 7 : 25 95 | 0 7 : 1 12)`. Note that both the `timeint()` option and `timeintknots()` option can be used with the `timeintbknots()` option and that knots in `timeintbknots()` should be specified on the untransformed timescale.

`noconstant` suppresses the constant term (intercept) in the model.

`nodes(`*#*`)` specifies the number of nodes to be used in Gauss–Legendre quadrature numerical integration when calculating the estimated cumulative hazard function from the estimated hazard function. The default is `nodes(30)`. Changing the nodes may be useful if there are convergence problems. Too few nodes may result in a poor approximation involved in the numerical integration. Analyses should be performed to ensure the results are not sensitive to the number of nodes.

`noorthog` suppresses orthogonal transformation of spline variables.

`nohr` reports the coefficients instead of hazard ratios.

`verbose` details the process of the `stmt` command.

`from(`*matrix*`)` defines the parameter matrix of initial values to be used in maximum likelihood estimation. By default, `stmt` estimates initial hazard estimates by fitting a model on the log cumulative-hazard scale using the `stpm2` command.

`inith(`*varname*`)` defines initial hazard estimates to be used in maximum likelihood estimation. By default, `stmt` estimates initial hazard estimates by fitting a model on the log cumulative-hazard scale using the `stpm2` command.

maximize_options: `difficult`, `technique(`*algorithm_spec*`)`, `iterate(`#`)`, [`no`]`log`, `trace`, `gradient`, `showstep`, `hessian`, `shownrtolerance`, `tolerance(`#`)`, `ltolerance(`#`)`, `gtolerance(`#`)`, `nrtolerance(`#`)`, `nonrtolerance`, `from(`*init_specs*`)`; see [R] **Maximize**. These options are seldom used, but the `difficult` option may be useful if there are convergence problems.

4.2.1 Timescale-specific suboptions

`bknots(`*knots_list*`)` specifies the boundary knots for the timescale specified. *knots_list* is a two-element *numlist* giving the boundary knots. By default, these are located at the minimum and maximum of the uncensored event times. They are specified on the scale defined by `knscale()`.

`bknotstvc(`*knots_list*`)` gives the boundary knots for any restricted cubic splines created for the timescale when including an interaction between a covariate and the timescales. By default, these are the same as for the `bknots()` option. They are specified on the scale defined by `knscale()`. For example, `bknotstvc(x1 0.01 10 x2 0.01 8)` indicates that the boundary knots for the timescale should be at 0.01 and 10 when including an interaction with the variable `x1` and at 0.01 and 8 when including an interaction with the variable `x2`.

`df(`#`)` specifies the degrees of freedom for the restricted cubic spline function for the baseline function; the number of degrees of freedom does not include the constant term. # must be between 1 and 10. With 1 degree of freedom, a linear function is fit. The `knots()` option is not applicable if the `df()` option is specified. The knots are placed at equally spaced percentiles of the uncensored event times or log event times, depending on the `logtoff` option. For example, if suboption `df(5)` is specified in the `time1()` option with no `logtoff` suboption, knots are placed at the 20th, 40th, 60th, and 80th percentiles of the distribution of the uncensored log event times on the first timescale. Note that these are interior knots and there are also boundary knots placed at the minimum and maximum of the distribution of uncensored event times or log survival-times.

`dftvc(`*df_list*`)` specifies the degrees of freedom used when including interactions between the timescale and covariates in *df_list*. If there is more than one interaction and different degrees of freedom are requested, then the syntax `dftvc(x1:3 x2:2 1)` applies. This will use 3 degrees of freedom for covariate `x1`, 2 degrees of freedom for covariate `x2`, and 1 degree of freedom for all remaining interactions between covariates and the restricted cubic spline of the first timescale if used in the `time1()` option.

`indicator(`*varname*`)` specifies an indicator variable that expresses which observations have more than one timescale. The indicator variable should be coded 0 for those observations who did not have the second or third timescale and 1 for those who did. This could be useful when a secondary timescale is relevant only for a subset of the analysis population.

`knots(`*#* [*#* `...`]`)` specifies the knot locations for the timescale, as opposed to the locations set by the `df()` option. Note that the locations of the knots are placed on the scale defined by `knscale()`. However, the scale used by the restricted cubic splines function is always log time unless the `logtoff` option is specified. Default knot locations are determined by the `df()` option.

`knotstvc(`*knots_list*`)` defines the location of the interior knots when recalculating the restricted cubic splines for the specified timescale-covariate interaction. If different knots are required for different interactions, the option is specified as, for example, `knotstvc(x1 1 2 3 x2 1.5 3.5)`.

`knscale(`*scale*`)` sets the scale on which user-defined knots are specified for the specified timescale. `knscale(time)` denotes the original timescale, `knscale(log)` denotes the log timescale, and `knscale(centile)` specifies that the knots are taken to be centile positions in the distribution of the uncensored survival times or log survival-times depending on whether the `logtoff` option is specified. The default is `knscale(time)`.

`logtoff` smooths the specified timescale over time using restricted cubic splines. By default, smoothing is over log time.

`start(`*varname*`)` specifies the variable that is the difference between the timescale specified in `stset` and the timescale of interest. For example, if the first timescale of interest (`t1`) is time since diagnosis and the second timescale (`t2`) is attained age, attained age is equal to time since diagnosis plus the age at diagnosis; that is, `t2` = `t1` + age at diagnosis. Thus, in this example *varname* would be a variable that contains the age at diagnosis. This option is not for use when using `time1()`, because this timescale is specified when using the `stset` command.

`tvc(`*varlist*`)` gives the name of the variables that are to be included as part of an interaction with the specified timescale. Interactions between covariates and timescales are included by reformulating the timescale using a restricted cubic spline, as the user prefers. The degrees of freedom are specified using the `dftvc()` option.

5 The stmt postestimation command

5.1 Syntax

`predict` *newvar* [*if*] [*in*] `,` {`hazard`|`xb`} `time1var(`*varname*`)`
[`time2var(`*varname*`)` `time3var(`*varname*`)` `at(`*varname* *#* [*varname* *#* `...`]`)`
`ci` `nodes(`*#*`)` `per(`*#*`)` `zeros` `level(`*#*`)`]

5.2 Options

`hazard` predicts the hazard function. `hazard` or `xb` is required.

`xb` predicts the linear predictor, including the spline function. `hazard` or `xb` is required.

`time1var(`*varname*`)` specifies the variable in the dataset that defines the values of timescale 1 that the user wishes to predict over. `time1var()` is required.

`time2var(`*varname*`)` specifies the variable in the dataset that defines the values of timescale 2 that the user wishes to predict over.

`time3var(`*varname*`)` specifies the variable in the dataset that defines the values of timescale 3 that the user wishes to predict over.

`at(`*varname* # [*varname* # `...`]`)` requests that the covariates specified by the listed *varname*s be set to the listed # values. For example, `at(x1 1 x3 50)` would evaluate predictions at `x1` = `1` and `x3` = `50`. This is a useful way to obtain out-of-sample predictions. Note that if `at()` is used together with `zeros`, all covariates not listed in `at()` are set to zero. If `at()` is used without `zeros`, then all covariates not listed in `at()` are set to their sample values. See also `zeros`.

`ci` calculates a confidence interval for the requested statistics and stores the confidence limits in *newvar*`_lci` and *newvar*`_uci`.

`nodes(`#`)` specifies the number of nodes to be used when numerically integrating the estimated hazard function using Gauss–Legendre quadrature. Numerical integration is required when predicting the cumulative hazard and survival functions. The default is `nodes(30)`.

`per(`#`)` expresses hazard rates and differences in hazard rates per # person years.

`zeros` sets all covariates to zero (baseline prediction). For example, `predict s0, survival zeros` calculates the baseline survival function. See also `at()`.

`level(`#`)` specifies the confidence level as a percentage. The default is `level(95)` or as set by `set level`.

6 Example

The `stmt` command and predictions are illustrated here through an application to 2,982 patients diagnosed with breast cancer in Rotterdam. Patients are followed from primary surgery until death (due to any cause) in this illustrative example. Time from primary surgery is used as the first timescale, and attained age is then introduced as a second timescale. Grade of breast cancer is the exposure variable of interest; note that all women have a diagnosis of a grade 2 or grade 3 breast cancer.

6.1 One timescale

We first consider grade on the risk of breast cancer mortality while modeling time since surgery and ignoring attained age. The `stset` command is used as follows to define the timescale of interest:

```
. use http://www.stata-press.com/data/fpsaus/rott2
(Rotterdam breast cancer data, truncated at 10 years)
. stset os, failure(osi) scale(12)
Survival-time data settings

         Failure event: osi!=0 & osi<.
Observed time interval: (0, os]
     Exit on or before: failure
     Time for analysis: time/12
--------------------------------------------------------------------------
      2,982  total observations
          0  exclusions
--------------------------------------------------------------------------
      2,982  observations remaining, representing
      1,272  failures in single-record/single-failure data
 21,270.702  total analysis time at risk and under observation
                                                At risk from t =         0
                                     Earliest observed entry t =         0
                                          Last observed exit t =  19.28268
```

A flexible parametric survival model on the log-hazard scale as a function of one timescale using `stmt` can then be implemented as follows:

```
. stmt grade, time1(df(5)) nolog
Log likelihood = -3023.3924                              Number of obs = 2,982

-------------+----------------------------------------------------------------
             |  Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
xb           |
       grade |   1.659792   .1152621     7.30   0.000     1.448582    1.901798
-------------+----------------------------------------------------------------
rcs          |
     __t1_s1 |   .1130917   .0309638     3.65   0.000     .0524038    .1737795
     __t1_s2 |   .1179876   .0291052     4.05   0.000     .0609425    .1750326
     __t1_s3 |  -.1213544   .0299503    -4.05   0.000    -.1800559    -.062653
     __t1_s4 |  -.0914425   .0299644    -3.05   0.002    -.1501716   -.0327134
     __t1_s5 |   -.026898   .0318642    -0.84   0.399    -.0893507    .0355546
       _cons |   -4.12488   .1960625   -21.04   0.000    -4.509155   -3.740604
-------------+----------------------------------------------------------------
Note: Estimates are transformed only in the first equation to hazard ratios.
 Quadrature method: Gauss-Legendre with 30 nodes
```

This model includes a restricted cubic spline function that models the log time since surgery timescale with 5 degrees of freedom. We see from the model output that the mortality rate in those diagnosed with grade 3 breast cancer is 1.66 times that of those diagnosed with grade 2 breast cancer while using time since surgery as our timescale. We get similar results when fitting the equivalent model using `strcs` and when fitting the model on the log cumulative-hazard scale using `stpm2`.

```
. quietly estimates store stmt
. quietly stpm2 grade, df(5) scale(h)
. quietly estimates store stpm2
. quietly strcs grade, df(5)
. quietly estimates store strcs
. estimate table stmt strcs stpm2, eform keep(grade) se

─────────────┬──────────────────────────────────────
    Variable │    stmt        strcs        stpm2
─────────────┼──────────────────────────────────────
       grade │  1.6597924    1.6597925    1.6582902
             │   .1152621     .1152621    .11514289
─────────────┴──────────────────────────────────────
                                       Legend: b/se
```

We can additionally adjust for age at surgery as a time-fixed version of our second timescale, attained age, by including this as a covariate in our model, as shown in the following output. Here we see that the continuous effect of age at surgery is significant and slightly reduces the effect estimate of grade, indicating that now the mortality rate in those diagnosed with grade 3 breast cancer is 1.63 times that of those diagnosed with grade 2 breast cancer.

```
. quietly rename age agesurgery
. stmt grade agesurgery, time1(df(5)) nolog
Log likelihood = -2990.4001                             Number of obs = 2,982

─────────────┬────────────────────────────────────────────────────────────────
             │ Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
─────────────┼────────────────────────────────────────────────────────────────
xb           │
       grade │   1.629973    .113256     7.03   0.000     1.422447    1.867776
  agesurgery │   1.018382   .0022847     8.12   0.000     1.013914     1.02287
─────────────┼────────────────────────────────────────────────────────────────
rcs          │
     __t1_s1 │    .131522   .0310742     4.23   0.000     .0706177    .1924263
     __t1_s2 │   .1073886   .0291626     3.68   0.000      .050231    .1645461
     __t1_s3 │  -.1292379   .0300013    -4.31   0.000    -.1880393   -.0704364
     __t1_s4 │  -.0965942   .0300029    -3.22   0.001    -.1553989   -.0377895
     __t1_s5 │   -.030469   .0318703    -0.96   0.339    -.0929337    .0319957
       _cons │  -5.075964   .2296522   -22.10   0.000    -5.526074   -4.625854
─────────────┴────────────────────────────────────────────────────────────────
Note: Estimates are transformed only in the first equation to hazard ratios.
 Quadrature method: Gauss-Legendre with 30 nodes
```

If we instead decide to model the log hazard as a function of the second timescale, attained age, rather than age at surgery, we can do so by introducing the `start()` option of `stmt`. In `start()`, we specify the variable that defines the difference between the origins of the two timescales. In this example, the age at surgery is the difference between the two timescales. Thus, the following command fits a flexible parametric survival model modeling the log hazard as a function of the two timescales, where the attained age timescale is modeled using a restricted cubic spline with 3 degrees of freedom.

```
. stmt grade, time1(df(5)) time2(df(3) start(agesurgery)) nolog
Log likelihood = -2956.7515                         Number of obs = 2,982

             |  Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
xb           |
       grade |   1.609036   .1118154     6.84   0.000     1.404151    1.843816
-------------+----------------------------------------------------------------
rcs          |
     __t1_s1 |    .078144   .0316819     2.47   0.014     .0160487    .1402394
     __t1_s2 |   .1307119   .0291879     4.48   0.000     .0735047     .187919
     __t1_s3 |  -.1186399   .0300058    -3.95   0.000    -.1774502   -.0598297
     __t1_s4 |  -.0946102   .0300088    -3.15   0.002    -.1534263   -.0357941
     __t1_s5 |   -.030988   .0318674    -0.97   0.331    -.0934469    .0314709
     __t2_s1 |   .2271669   .0260366     8.72   0.000     .1761361    .2781977
     __t2_s2 |  -.2377882   .0251054    -9.47   0.000    -.2869938   -.1885825
     __t2_s3 |  -.0887876   .0266178    -3.34   0.001    -.1409575   -.0366178
       _cons |  -4.048158   .1961963   -20.63   0.000    -4.432695    -3.66362
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to hazard ratios.
 Quadrature method: Gauss-Legendre with 30 nodes
. quietly estimates store stmt_2ts
```

When modeling as a function of two timescales, the mortality rate for those with grade 3 breast cancer is 1.61 times the rate in those with grade 2 breast cancer. We can also include an interaction between the timescale and the grade covariate if we wish, allowing the hazard ratio for grade to vary over the timescale. Here we do this for the time-since-surgery timescale.

```
. stmt grade, time1(df(5) tvc(grade) dftvc(2)) time2(df(3) start(agesurgery))
> nolog
Log likelihood = -2953.3856                         Number of obs = 2,982

               |  Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
---------------+----------------------------------------------------------------
xb             |
         grade |   1.514219   .1233225     5.09   0.000     1.290816    1.776287
---------------+----------------------------------------------------------------
rcs            |
       __t1_s1 |   .5111851   .2011859     2.54   0.011      .116868    .9055022
       __t1_s2 |   .3295628   .2191507     1.50   0.133    -.0999648    .7590903
       __t1_s3 |   -.098055   .0353463    -2.77   0.006    -.1673326   -.0287775
       __t1_s4 |   -.091684   .0300071    -3.06   0.002    -.1504969   -.0328711
       __t1_s5 |  -.0309256   .0319089    -0.97   0.332     -.093466    .0316147
       __t2_s1 |   .2274068   .0260295     8.74   0.000     .1763899    .2784236
       __t2_s2 |  -.2386316   .0251059    -9.51   0.000    -.2878382    -.189425
       __t2_s3 |  -.0892667   .0266396    -3.35   0.001    -.1414793   -.0370541
__t1_s_grade1  |  -.1539901   .0712874    -2.16   0.031    -.2937108   -.0142695
__t1_s_grade2  |   -.068185   .0770798    -0.88   0.376    -.2192586    .0828886
         _cons |  -3.884596   .2262782   -17.17   0.000    -4.328093   -3.441099
--------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to hazard ratios.
 Quadrature method: Gauss-Legendre with 30 nodes
```

We may also wish to include an interaction between the two timescale functions. We use the `timeint()` option to define the two timescales and the degrees of freedom that will be used for the interaction.

```
. stmt grade, time1(df(5)) time2(df(3) start(agesurgery)) timeint(t1:t2 2:2)
> nolog
Log likelihood = -2942.5024                            Number of obs = 2,982

------------------------------------------------------------------------------
             |  Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
xb           |
       grade |   1.604792   .1115284     6.81   0.000     1.400434    1.838971
-------------+----------------------------------------------------------------
rcs          |
      __t1_s1|   .0484406   .0333889     1.45   0.147    -.0170004    .1138817
      __t1_s2|   .1548197   .0317914     4.87   0.000     .0925096    .2171297
      __t1_s3|  -.1052611    .030662    -3.43   0.001    -.1653575   -.0451646
      __t1_s4|  -.0690362    .030844    -2.24   0.025    -.1294893    -.008583
      __t1_s5|  -.0299301   .0317676    -0.94   0.346    -.0921934    .0323332
      __t2_s1|   .2719147   .0418266     6.50   0.000     .1899361    .3538933
      __t2_s2|  -.2557761   .0408903    -6.26   0.000    -.3359197   -.1756325
      __t2_s3|  -.0730147   .0279576    -2.61   0.009    -.1278106   -.0182188
  __t1_t2_s11|     .02997   .0318829     0.94   0.347    -.0325193    .0924594
  __t1_t2_s12|  -.0339458   .0313896    -1.08   0.280    -.0954683    .0275767
  __t1_t2_s21|  -.1274015   .0295475    -4.31   0.000    -.1853136   -.0694895
  __t1_t2_s22|   .0200843     .02825     0.71   0.477    -.0352846    .0754532
        _cons|  -4.073179   .1967425   -20.70   0.000    -4.458788   -3.687571
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to hazard ratios.
 Quadrature method: Gauss-Legendre with 30 nodes
. quietly estimates store stmt_timeint
```

Here time since surgery is modeled using a spline with 5 degrees of freedom and attained age with 3 degrees of freedom. The interaction terms are estimated by re-creating restricted cubic splines on each timescale with 2 degrees of freedom each (specified in the `timeint()` option) and multiplying these together. This model indicates that the mortality rate in those with grade 3 breast cancer is 1.60 times that of individuals with grade 2 breast cancer once we include these functions in the model.

Likelihood-ratio tests can be performed when models are nested. Here we assess whether the timescale interaction is significant by using the `lrtest` command and see that the model including the timescale interaction is significantly better than the proportional hazards model with two timescales.

```
. lrtest stmt_2ts stmt_timeint
Likelihood-ratio test
Assumption: stmt_2ts nested within stmt_timeint
 LR chi2(4) =  28.50
Prob > chi2 = 0.0000
```

Hazard rates from the models fit with `stmt` can be predicted as described below. The user must first create variables that represent the timescales to be predicted over; these created variables are then fed into the `predict` command. For example, should we wish to predict the hazard function for those diagnosed with grade 3 breast cancer from start to 10 years since surgery (timescale 1) for those at 50 years on the attained age timescale (timescale 2), we can create variables `time1` and `time2` and feed them into the `predict` command:

```
. range time1 0 10 100
(2,882 missing values generated)
. generate time2=50
. predict h, hazard time1var(time1) time2var(time2) at(grade 3)
(2,883 missing values generated)
```

We can also predict the hazard rate across both timescales simultaneously via the creation of the timescale variables that are fed into the `predict` command. For example, we can make predictions similar to those above but for every combination of time since surgery and attained age as follows. We start by creating a temporary dataset containing every combination of timescale 1 (time since diagnosis, from 0.2 to 15 years in 0.2 yearly steps) and timescale 2 (attained age from 40 to 70 years in 301 steps using the `range` command).

```
. capture drop time1 time2
. preserve
. forvalues j=0.2(0.2)15 {
  2.         quietly clear
  3.         quietly set obs 301
  4.         quietly generate time1 = `j'
  5.         quietly range time2 40 70 301
  6.         quietly tempfile temppred`n'
  7.         quietly save `temppred`n''
  8.         local datalist `datalist' `temppred`n''
  9.         local n=`n'+1
 10. }
. clear
. quietly set obs 0
. quietly append using `datalist'
. quietly tempfile timedata
. quietly save `timedata'
. restore
```

Below, the first 15 rows of this temporary dataset are displayed.

```
. list time1 time2 in 1/15

     ┌───────────────┐
     │ time1   time2 │
     ├───────────────┤
  1. │    .2      40 │
  2. │    .2    40.1 │
  3. │    .2    40.2 │
  4. │    .2    40.3 │
  5. │    .2    40.4 │
     ├───────────────┤
  6. │    .2    40.5 │
  7. │    .2    40.6 │
  8. │    .2    40.7 │
  9. │    .2    40.8 │
 10. │    .2    40.9 │
     ├───────────────┤
 11. │    .2      41 │
 12. │    .2    41.1 │
 13. │    .2    41.2 │
 14. │    .2    41.3 │
 15. │    .2    41.4 │
     └───────────────┘
```

This temporary dataset is then merged into our original dataset, and these values of the timescales are used to predict the mortality rate using the `hazard` option like before (note that the `ageatsurgery` variable is not used for our predictions but rather used later for visualization of these hazard rates):

```
. merge 1:1 _n using `timedata', nogen
  (output omitted)
. generate ageatsurgery=round(time2-time1, 0.1)
. predict h2, hazard time1var(time1) time2var(time2) at(grade 3) ci
note: confidence intervals calculated using Z critical values.
```

Using the predicted values, we can plot different combinations of the mortality rate to address different questions. For example, if we were interested in the breast cancer mortality rate for those diagnosed with grade 3 breast cancer over attained age for different ages at surgery, we can use the following code and present the rates as in figure 1.

```
. twoway (line h2 time2 if ageatsurgery==40, sort)
>        (line h2 time2 if ageatsurgery==50, sort)
>        (line h2 time2 if ageatsurgery==60, sort),
>        legend(order(1 "40 years" 2 "50 years" 3 "60 years" ) rows(1))
>        ylabel(, format(%3.2f) angle(h))
>        scheme(sj)
>        xtitle("Attained age (years)")
>        ytitle("Mortality rate (per person-year)")
>        name(overts2_agesurgery, replace)
```

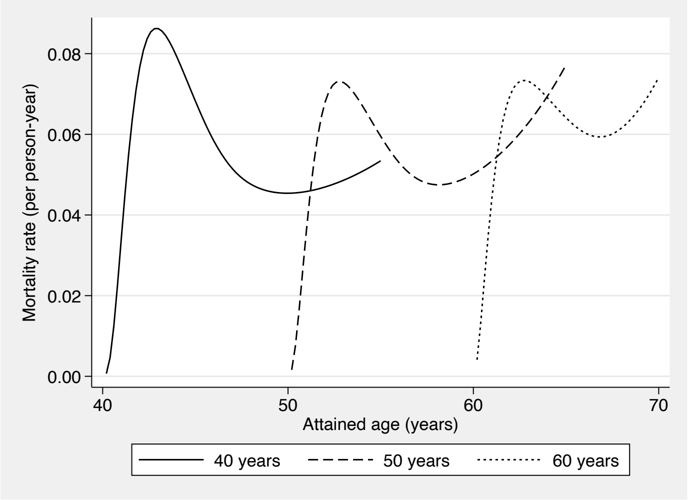
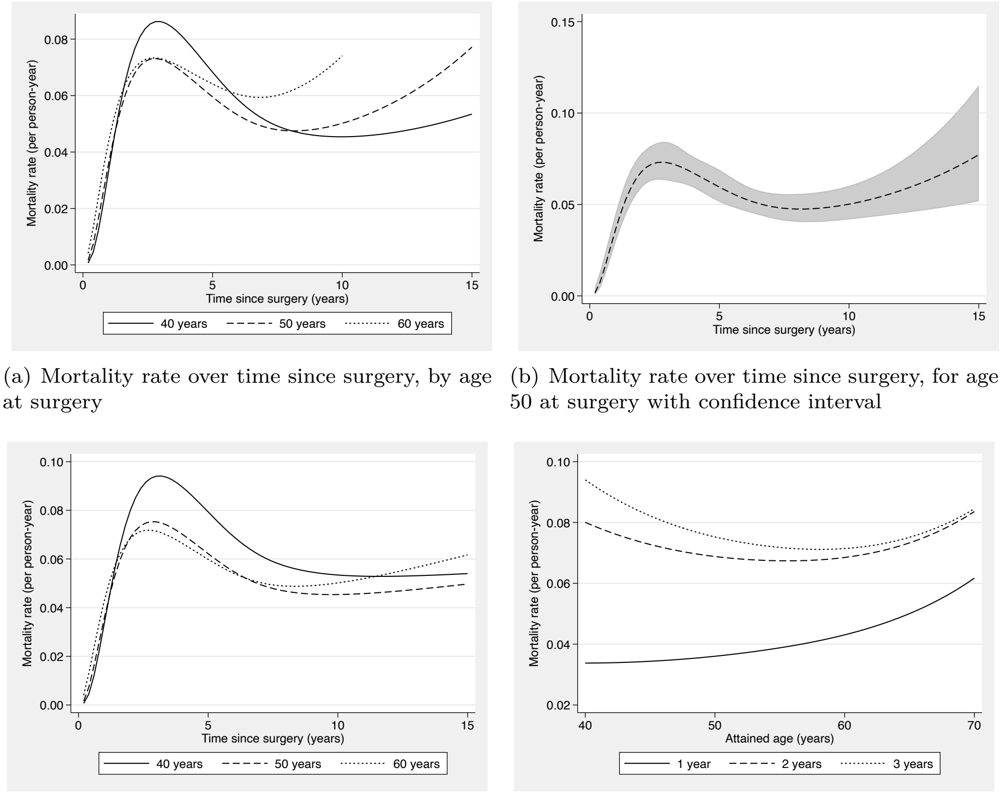


Figure 1. Mortality rate of grade 3 breast cancer presented across attained age for different ages at surgery

Other ways of presenting the mortality rate for grade 3 breast cancer are shown in figure 2.



(a) Mortality rate over time since surgery, by age at surgery

(b) Mortality rate over time since surgery, for age 50 at surgery with confidence interval

(c) Mortality rate over time since surgery, by different attained ages

(d) Mortality rate over attained age, by time since surgery

Figure 2. Other mortality prediction presentations available in `stmt`

We can also predict hazard ratios using the `predictnl` command. Below, we present code that estimates and plots the hazard ratio comparing patients diagnosed with grade 3 breast cancer with those with grade 2 breast cancer for the timescales we previously created. We first start by simplifying the model back to one with no timescale interactions to simplify presentation of results. Figure 3 shows the predicted hazard ratio over time since surgery, illustrating that the relative mortality rate in those with grade 3, compared with grade 2, is higher early on after surgery and after 1–2 years remains around 1.5 (at 5 years, hazard ratio = 1.48, 95% confidence interval [1.26, 1.75]). Note that when predicting the hazard ratio, we estimate on the log-hazard scale and then back-transform to the hazard scale.

```
. stmt grade, time1(df(5) tvc(grade) dftvc(2)) time2(df(3) start(agesurgery))
Iteration 0:   log likelihood = -3021.6264
Iteration 1:   log likelihood = -2969.4752
Iteration 2:   log likelihood = -2953.4512
Iteration 3:   log likelihood = -2953.3856
Iteration 4:   log likelihood = -2953.3856
Log likelihood = -2953.3856                             Number of obs = 2,982
------------------------------------------------------------------------------
              | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
--------------+---------------------------------------------------------------
xb            |
        grade |   1.514219   .1233225     5.09   0.000     1.290816    1.776287
--------------+---------------------------------------------------------------
rcs           |
       __t1_s1 |   .5111851   .2011859     2.54   0.011      .116868    .9055022
       __t1_s2 |   .3295628   .2191507     1.50   0.133    -.0999648    .7590903
       __t1_s3 |   -.098055   .0353463    -2.77   0.006    -.1673326   -.0287775
       __t1_s4 |   -.091684   .0300071    -3.06   0.002    -.1504969   -.0328711
       __t1_s5 |  -.0309256   .0319089    -0.97   0.332     -.093466    .0316147
       __t2_s1 |   .2274068   .0260295     8.74   0.000     .1763899    .2784236
       __t2_s2 |  -.2386316   .0251059    -9.51   0.000    -.2878382    -.189425
       __t2_s3 |  -.0892667   .0266396    -3.35   0.001    -.1414793   -.0370541
__t1_s_grade1 |  -.1539901   .0712874    -2.16   0.031    -.2937108   -.0142695
__t1_s_grade2 |   -.068185   .0770798    -0.88   0.376    -.2192586    .0828886
        _cons |  -3.884596   .2262782   -17.17   0.000    -4.328093   -3.441099
------------------------------------------------------------------------------
Note: Estimates are transformed only in the first equation to hazard ratios.
 Quadrature method: Gauss-Legendre with 30 nodes
. predictnl lnhr= ln(predict(h time1var(time1) time2var(time2) at(grade 3)))
> - ln(predict(h time1var(time1) time2var(time2) at(grade 2))),
> ci(lnhr_lci lnhr_uci)
note: confidence intervals calculated using Z critical values.
. generate hr=exp(lnhr)
. generate hr_lci=exp(lnhr_lci)
. generate hr_uci=exp(lnhr_uci)
. twoway (rarea hr_lci hr_uci time1 if time1>=0.25, sort lcolor(gs12) fcolor(gs12))
>        (line hr time1 if time1>=0.25, sort),
>        legend(off)
>        ylabel(, format(%2.1f) angle(h))
>        scheme(sj)
>        xtitle("Time since surgery (years)")
>        ytitle("Hazard ratio")
>        name(hr_grade, replace)
```

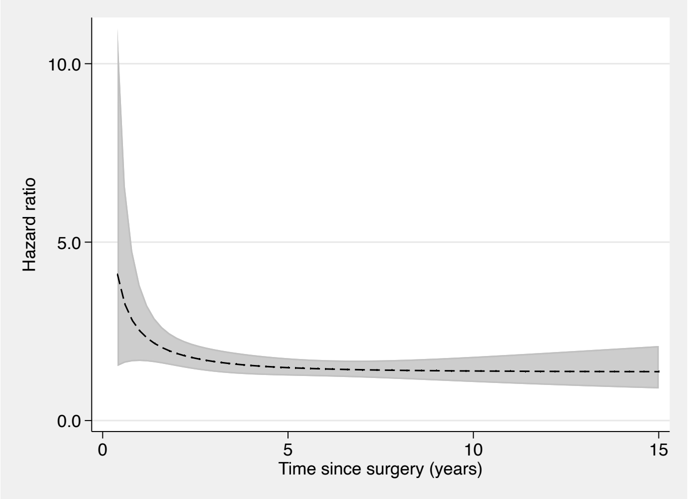


Figure 3. Hazard ratio (grade 3 versus grade 2 breast cancer) presented over time since surgery. Note that values are presented after 0.25 years (time since surgery).

7 Conclusion

We illustrated how flexible parametric survival models on the log-hazard scale can be used to model multiple timescales and how to fit these models in Stata using the stmt command. The stmt command is a user-friendly tool that comes with a handy postestimation command that can produce practical predictions to illustrate different effects on different timescales.

8 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 22-3
. net install st0688     (to install program files, if available)
. net get st0688         (to install ancillary files, if available)
```

stmt can also be downloaded from the Statistical Software Components Archive by typing

```
. ssc install stmt
```

9 References

Albrektsen, G., I. Heuch, S. Thoresen, and G. Kvåle. 2006. Clinical stage of breast cancer by parity, age at birth, and time since birth: A progressive effect of pregnancy hormones? *Cancer Epidemiology Biomarkers Prevention* 15: 65–69. https://doi.org/10.1158/1055-9965.EPI-05-0634.

Batyrbekova, N., H. Bower, P. W. Dickman, R. Szulkin, P. C. Lambert, and T. M.-L. Andersson. Forthcoming. Potential bias introduced by not including multiple time-scales in survival analysis: A simulation study. *Communications in Statistics—Simulation and Computation*. https://doi.org/10.1080/03610918.2022.2038626.

Bower, H., M. J. Crowther, and P. C. Lambert. 2016. strcs: A command for fitting flexible parametric survival models on the log-hazard scale. *Stata Journal* 16: 989–1012. https://doi.org/10.1177/1536867X1601600410.

Carstensen, B. 2006. Demography and epidemiology: Practical use of the Lexis diagram in the computer age, or: Who needs the Cox-model anyway? Technical Report 06.2, Department of Biostatistics, University of Copenhagen.

———. 2007. Age–period–cohort models for the Lexis diagram. *Statistics in Medicine* 26: 3018–3045. https://doi.org/10.1002/sim.2764.

Crowther, M. J., and P. C. Lambert. 2014. A general framework for parametric survival analysis. *Statistics in Medicine* 33: 5280–5297. https://doi.org/10.1002/sim.6300.

Danardono, D. 2005. Multiple time scales and longitudinal measurements in event history analysis. PhD thesis, Umeå University. http://umu.diva-portal.org/smash/get/diva2:143422/FULLTEXT01.pdf.

Duchesne, T., and J. Lawless. 2000. Alternative time scales and failure time models. *Lifetime Data Analysis* 6: 157–179. https://doi.org/10.1023/A:1009616111968.

Efron, B. 2002. The two-way proportional hazards model. *Journal of the Royal Statistical Society, Series B* 64: 899–909. https://doi.org/10.1111/1467-9868.00368.

Iacobelli, S., and B. Carstensen. 2013. Multiple time scales in multi-state models. *Statistics in Medicine* 32: 5315–5327. https://doi.org/10.1002/sim.5976.

Lambert, P. 2008. rcsgen: Stata module to generate restricted cubic splines and their derivatives. Statistical Software Components S456986, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s456986.html.

———. 2010. stpm2: Stata module to estimate flexible parametric survival models. Statistical Software Components S457128, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s457128.html.

Royston, P., and P. C. Lambert. 2011. *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model.* College Station, TX: Stata Press.

Rutherford, M. J., M. J. Crowther, and P. C. Lambert. 2015. The use of restricted cubic splines to approximate complex hazard functions in the analysis of time-to-event data: A simulation study. *Journal of Statistical Computation and Simulation* 85: 777–793. https://doi.org/10.1080/00949655.2013.845890.

Stoer, J., and R. Bulirsch. 2002. *Introduction to Numerical Analysis*. 3rd ed. New York: Springer. https://doi.org/10.1007/978-0-387-21738-3.

Wolkewitz, M., B. S. Cooper, M. Palomar-Martinez, F. Alvarez-Lerma, P. Olaechea-Astigarraga, A. G. Barnett, and M. Schumacher. 2016. Multiple time scales in modeling the incidence of infections acquired in intensive care units. *BMC Medical Research Methodology* 16: 116. https://doi.org/10.1186/s12874-016-0199-y.

About the authors

Hannah Bower is a research coordinator and biostatistician at the Department of Medicine Solna, Karolinska Institutet, Stockholm, Sweden.

Therese M.-L. Andersson is a senior lecturer in the Department of Medical Epidemiology and Biostatistics at Karolinska Institutet, Stockholm, Sweden.

Michael J. Crowther is a biostatistician in the Department of Medical Epidemiology and Biostatistics at Karolinska Institutet, Stockholm, Sweden.

Paul C. Lambert is a professor of biostatistics at the University of Leicester in Leicester, UK. Paul has a long-term secondment arrangement with the Department of Medical Epidemiology and Biostatistics at Karolinska Institutet.