The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–26

stpm2cr: A flexible parametric competing risks model using a direct likelihood approach for the cause-specific cumulative incidence function

Sarwar Islam Mozumder
Department of Health Sciences
University of Leicester
Leicester, UK
si113@le.ac.uk

Mark J. Rutherford
Department of Health Sciences
University of Leicester
Leicester, UK
mark.rutherford@le.ac.uk

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

Abstract. In a competing risks analysis, interest lies in the cause-specific cumulative incidence function (CIF) which is usually obtained in a modelling framework by either (1) transforming on all of the cause-specific hazard (CSH) or (2) through its direct relationship with the subdistribution hazard (SDH) function. We expand on current competing risks methodology from within the flexible parametric survival modelling framework (FPM) and focus on approach (2). This models all cause-specific CIFs simultaneously and is more useful when prognostic related questions are to be answered. We propose the direct FPM approach for the cause-specific CIF which models the (log-cumulative) baseline hazard without the requirement of numerical integration leading to benefits in computational time. It is also easy to make out-of-sample predictions to estimate more useful measures and alternative link functions can be incorporated, for example, the logit link. To implement the methods, a new estimation command, `stpm2cr`, is introduced and useful predictions from the model are demonstrated through an illustrative Melanoma dataset.

Keywords: st0001, stpm2cr, survival analysis, competing risks, flexible parametric models, subdistribution hazard, cumulative incidence function

1 Introduction

In competing risks, the cause-specific cumulative incidence function (CIF), which is the probability of failure from an event in the presence of other competing events, is considered. From within the modelling framework this is usually obtained by either

(1) estimating all the cause-specific hazard (CSH) functions, or (2) transforming using a direct relationship with the subdistribution hazard (SDH) function for the cause of interest. There are a number of different tools available in Stata that allow us to estimate the cause-specific CIF. An empirical, non-parametric estimate of the cause-specific CIF can be obtained using the user-written command `stcompet` which applies the Aalen-Johansen approach Coviello and Boggess (2004).

Alternatively, we can fit regression models on either the CSH or SDH scale, the choice of which relates to the research question to be answered (Sapir-Pichhadze et al. 2016; Noordzij et al. 2013; Koller et al. 2012). CSH regression models can be fitted from within a semi-parametric approach using a typical Cox model or from within a flexible parametric modelling framework, using the user-written post-estimation command, `stpm2cif`. This command is used with an expanded dataset where each patient has a row for each cause and after fitting a cause-specific flexible parametric survival model (FPM) with `stpm2` to model all causes (Hinchliffe and Lambert 2013; Lambert and Royston 2009; Lambert et al. 2011; Royston and Parmar 2002).

The most popularly applied method for modelling covariate effects on the cause-specific CIF is the Fine & Gray model (Fine and Gray 1999) and is available through the `stcrreg` command. However, this approach only allows us to model one event individually using the partial-likelihood and we must fit separate models for each competing event if we want to understand the overall impact of a covariate on risk.

Competing risks models can also be fit using the user-written `stcrprep` command which restructures the data and calculates the appropriate weights. Standard Stata survival analysis commands can then be used to fit models more computationally efficiently such as the Fine & Gray model and parametric models for the cause-specific CIF (Lambert et al. 2016 (submitted).

We introduce the use of parametric methods using the full-likelihood as smooth estimates can be obtained for the baseline cause-specific CIF or SDH for a particular cause which can easily extend to incorporate non-proportional SDHs. Fitting parametric models for the cause-specific CIF in this way is computationally quicker than fitting models with `stcrprep` since no numerical integration or data restructure is required. An additional advantage of these models is that we are able to model all cause-specific CIFs simultaneously and covariate effects are modelled on all competing causes. Jeong and Fine (2006) investigated a direct parametric inference approach and define a likelihood which allows us to model all the cause-specific CIFs simultaneously. We extend this approach to FPMs where it is easy to model time-dependent effects and obtain useful out-of-sample predictions.

Others have also proposed modelling the SDH under alternative link functions. For example, Gerds et al. (2012) proposes the proportional log-odds model for the cause-specific CIF which offers an alternative interpretation. However, the interpretation is not as simple as it is when modelling a single event and suffers from similar issues in interpretation as in the complementary log-log link function. Incorporating such alternative link functions on the cause-specific CIF is also easy to implement using the approach outlined in this paper.

The remaining content of this paper is structured as follows. In Section 2, we begin by introducing the methods for direct inference on the cause-specific CIF under a FPM framework. Section 3 outlines the syntax of `stpm2cr` which fits the models introduced in Section 2 and in Section 4, syntax for postestimation using `predict` after fitting models with `stpm2cr` is described. This is followed by some illustrative examples in Section 2. Finally, the paper is concluded with some discussion on the approach including limitations and potential extensions.

2 Methods

Let T be the time to event for any K competing causes $k = 1, \cdots, K$ and D denote the type of event, where $D = 1, \cdots, K$. Here, we consider the events to be death from different causes and so the cause-specific CIF, $F_k(t)$, is the probability of dying from a particular cause, $D = k$, by time t whilst also being at risk of dying from other causes (Putter et al. 2007),

$$F_k(t) = P(T \leq t, D = k) \tag{1}$$

The all-cause CIF, $F(t)$, which is the probability of dying from any of the K causes by time t, is the sum of all K cause-specific CIFs, $F_k(t)$, and can also be expressed as the complement of the overall survival function, $S(t)$,

$$F(t) = P(T \leq t) = \sum_{j=1}^{K} F_j(t) = 1 - S(t) \tag{2}$$

2.1 Cause-specific hazard function

The cause-specific CIF, $F_k(t)$, can be expressed as a function of either the CSH functions for all K causes or the SDH for cause k. The CSH function, $h_k^{cs}(t)$ gives the instantaneous mortality rate from a particular cause k given that the patient is still alive at time t.

$$h_k^{cs}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t)}{\Delta t} \tag{3}$$

The cause-specific CIF can be expressed as a function of the CSHs for all K causes such that,

$$F_k(t) = \int_0^t \left(\exp \left[- \int_0^t \sum_{j=1}^{K} h_j^{cs}(u) du \right] \right) h_k^{cs}(u) du \tag{4}$$

Note here that the leading term within the integral gives the overall survival function, $S(t)$,

$$S(t) = \exp\left[-\int_0^t \sum_{j=1}^{K} h_j^{cs}(u)du\right] \tag{5}$$

2.2 Subdistribution hazard function

Gray (1988) introduces the SDH for cause k, $h_k^{sd}(t)$, which gives a direct relationship with the cause-specific CIF. This has the following mathematical formulation,

$$h_k^{sd}(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t, D = k | T > t \cup (T \leq t \cap D \neq k)}{\Delta t} \tag{6}$$

$$= \frac{\frac{d}{dt}\left[F_k(t)\right]}{1 - F_k(t)} = -\frac{d\left[\ln\left(1 - F_k(t)\right)\right]}{dt} \tag{7}$$

and is interpreted as the instantaneous rate of failure at time t from cause k amongst those who are still alive, or have died from any of the other $K-1$ competing causes excluding cause k. The SDH rate is not a conventional epidemiological rate due to the risk-set (see Lau et al. (2009)) and should not be interpreted as a standard hazard rate.

The cause-specific CIF can be expressed directly in terms of the SDH function for cause k using standard survival relationships along with the cumulative SDH for cause k, $H_k^{sd}(t)$,

$$F_k(t) = 1 - \exp\left[-H_k^{sd}(u)\right] \qquad \text{and} \qquad H_k^{sd}(t) = \int_0^t h_k^{sd}(u)du \tag{8}$$

Using the SDH functions for all K causes, we can also obtain the CSH functions, $h_k^{cs}(t)$, for all K causes (Beyersmann and Schumacher 2007),

$$h_k^{cs}(t) = h_k^{sd}(t)\left[1 + \frac{\left[\sum_{j=1}^{K} F_j(z)\right] - F_k(t)}{1 - \sum_{j=1}^{K} F_j(t)}\right] \tag{9}$$

2.3 Regression modelling

The most common model for the SDH for cause k is the Fine & Gray model (Fine and Gray 1999), which is expressed in a similar way to the cause-specific Cox PH model in that it assumes proportionality of covariate effects on the SDH scale,

$$h_k^{sd}(t|\mathbf{x}) = h_{0,k}^{sd}(t)\exp\left[\mathbf{x}_k\boldsymbol{\beta}_k^{sd}\right] \tag{10}$$

where $\boldsymbol{\beta}_k^{sd}$ are log-SDH ratios (SHR) for cause k. The SHR, $\exp\left(\boldsymbol{\beta}_k^{sd}\right)$ is interpreted as the association on the effect of a covariate on risk (refer to Wolbers et al. (2014) for more details on interpretation). We focus on implementing and extending the SDH regression model in Equation **??** from within the FPM approach.

2.4 Likelihood estimation

Jeong and Fine (2006) showed that we can simultaneously fit parametric models that directly estimate covariate effects on the cause-specific CIF for all k causes, $F_k(t|\mathbf{x}_k)$ $(k = 1, \cdots, K)$, without the requirement of indirect specification through the CSHs. Hence, for an observable failure time t_i, with independent right censoring, for each individual $i = 1, \cdots, N$, the likelihood for direct inference on the cause-specific CIF is,

$$L = \prod_{i=1}^{N} \left[\prod_{j=1}^{K} \left[h_j^{sd}(t_i|\mathbf{x_j})(1 - F_j(t_i)) \right]^{\delta_{ij}} \left[1 - \sum_{j=1}^{K} F_j(t_i|\mathbf{x}_j) \right]^{1-\sum_{j=1}^{K} \delta_{ij}} \right] \tag{11}$$

where the censoring indicator, δ_{ik}, tell us whether an individual died from any cause k ($\delta_{ik} = 1$), or not ($\delta_{ik} = 0$). Note here, however, that, the cause-specific CIF, $F_k(t)$, in Equation **??** is not a proper cumulative distribution function and is instead referred to as a subdistribution function since $\lim_{t\to\infty} F_k(t) \neq 1$ (Andersen et al. 2012).

2.5 Flexible parametric regression on the cause-specific cumulative incidence function

Using the likelihood in Equation 11, a parametric survival model can be fitted simultaneously for all K cause-specific CIFs. We apply the likelihood to the FPM approach described by Royston and Parmar (2002) and extend on this using restricted cubic splines, $s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k)$, with $M - 1$ degrees of freedom where s_k is a restricted cubic spline function for cause k on log-time and consists of a vector of M knots, $\mathbf{m}$, a vector of $M - 1$ parameters, $\boldsymbol{\gamma}$ and covariates $\mathbf{x}_k$ (Durrleman and Simon 1989). The following model can be specified through a general link function, $g(\cdot)$, for each of the $k = 1, \cdots, K$ cause-specific CIF with covariates, $\mathbf{x}_k$,

$$\begin{aligned} g(F_k(t|\mathbf{x}_{ik})) &= s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_k) + \mathbf{x_k}\boldsymbol{\beta}_k & (12) \\ &= \gamma_{0k} + \gamma_{1k} z_{1k} + \cdots + \gamma_{(M-1)k} z_{(M-1)k} + \mathbf{x}_k\boldsymbol{\beta}_k & (13) \end{aligned}$$

Where $z_{1k}, \cdots, z_{(M-1)k}$ are the basis functions of the restricted cubic splines and are defined as follows:

$$z_{1k} = \ln(t) \tag{14}$$

$$z_{jk} = (\ln(t) - m_{jk})_+^3 - \phi_{jk}(\ln(t) - m_{1k})_+^3 - (1 - \phi_{jk})(\ln(t) - m_{Mk})_+^3, \quad j = 2, \cdots, M-1$$

where,

$$\phi_{jk} = \frac{m_{Mk} - m_{jk}}{m_{Mk} - m_{1k}} \tag{15}$$

and

$$(u)_+ = \begin{cases} u, & \text{if } u < 0 \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Through the general link function, $g(\cdot)$, for the cause-specific CIF, $F_k(t)$, in Equation 12, are able to apply similar transformations described in Royston and Parmar (2002) for the survival function. Lambert et. al. (submitted) offers more details on the various link functions available for the cause-specific CIF, but here we only introduce the complementary log-log (cloglog) and logit link function (see Table 1).

Table 1: Common transformations on the general link function for the cause-specific CIF

Parameters	Link Function	Link Name
log-subdistribution hazard ratios	$\ln[-\ln(1 - F_k(t\|\mathbf{x}_k))]$	cloglog
log-odd ratios	$\frac{F_k(t\|\mathbf{x}_k)}{1-F_k(t\|\mathbf{x}_k)}$	logit

2.6 Time-dependent effects

To relax the proportionality assumption, interactions are fitted between the associated covariates and the spline function for log-time. This allows us to introduce a new set of knots, $\mathbf{m}_{ek}$, which represent the eth time-dependent effect for cause k with associated parameters $\boldsymbol{\alpha}_{ek}$. If there are $e = 1, \cdots, E$ time-dependent effects, then we can extend the model in Equation 12 to,

$$\eta_k(t) = s_k(\ln(t); \boldsymbol{\gamma}_k, \mathbf{m}_{0k}) + \mathbf{x}_k\boldsymbol{\beta}_k + \sum_{l=1}^{E} s_k(\ln(t); \boldsymbol{\alpha}_{lk}, \mathbf{m}_{lk})x_{lk} \tag{17}$$

In this approach, the spline function for different time-dependent effects can be different and usually requires fewer knots for the baseline spline function. This is an

extension on the original approach proposed by Royston and Parmar (2002). As all K causes are modelled, it is also possible to specify different time-dependent effects for the each of the k cause-specific FPM regression model.

2.7 Delayed entry

`stpm2cr` can also model left-truncated data or data with delayed entry. This is when subjects are considered to be at risk some time after $t = 0$.

2.8 Cure models

Andersson et al. (2011) proposed a method that allows estimation of the cure proportion in a relative survival FPM framework. In the competing risks scenario, this would occur in a situation where the cause-specific CIF is constant after a certain point in time t. Hence, by adapting the approach described by Andersson et al. (2011), we can estimate the cure proportion from within a flexible parametric model for the cause-specific CIF specified in Section 2.5 by forcing the log cumulative SDH to plateau after the last knot. This involves an adjustment to the way the spline variables are calculated so that the cause-specific CIF is forced to plateau (see Andersson et al. (2011) for more details). Since the SDH function for cause k on which we assume cure needs to be evaluated whilst simultaneously modelling all other causes, the final knot must be specified after the final observed time of death which has been set at the 110^{th} percentile of log-time. Applying the methods in Andersson et al. (2011) and the above adjustment to a specific cause $k = c$, we can fit a flexible parametric cure model with a complementary log-log link for a cause-specific CIF such that,

$$F_c(t|\mathbf{x}_c) = 1 - (1 - \pi_c)^{\exp\left[\gamma_{2c}z_{2c}+\cdots+\gamma_{(M-1)c}z_{(M-1)c}+\sum_{i=1}^{E} s_c(\ln(t);\boldsymbol{\alpha}_{ic},\mathbf{m}_{ic})\mathbf{x}_{ic}\right]} \tag{18}$$

$$1 - \pi_c = 1 - \exp\left(-\exp(\gamma_{0c} + \mathbf{x}_c\boldsymbol{\beta}_c)\right) \tag{19}$$

Therefore, the parameters, γ_{0c} and $\boldsymbol{\beta}_c$ are used to estimate the cure proportion for cause $k = c$. Here, we also implement a constraint on the linear spline, γ_{1c}, such that it is equal 0.

To fit a cure model, a plateau needs to be observed in the "raw" data for the cause-specific CIF on which we wish to model cure. This is usually done for a single relevant cause, particularly the event of interest.

3 Syntax

`stpm2cr` [*equation1*][*equation2*]...[*equationN*] [*if*] [*in*] , `events(`*varname*`)` [
`censvalue(`#`)` `cause(`*numlist*`)` `level(`#`)` `alleq` `noorthog` `eform` `oldest`
`mlmethod(`*string*`)` `lininit` *maximise_options*]

Where *equation1*, *equation2*, ..., *equationN* are the equations for each competing event. Note that at least two equations must be specified. The syntax of each equation is:

causename:[*varlist*], scale(*scalename*) [df(#) knots(*numlist*) tvc(*varlist*) dftvc(*df_list*) knotstvc(*numlist*) bknots(*knotslist*) bknotstvc(*numlist*) noconstant cure]

You must `stset` your data before using `stpm2cr`; see [ST] **stset**. All events must be specified in the `failure` option of `stset`.

3.1 Main Options

Model

`events(`*varname*`)` specifies the *varname* that contains the indicators for each competing event failure.

`cause(`*numlist*`)` specifies the indicator value(s) for the competing events specified in `events()`. The indicators specified in *numlist* must be listed in the same order of the equations *equation1*, *equation2*, ..., *equationN*.

`censvalue(#)` specifies the indicator value(s) in `events()` for individuals that are censored; The default is `censvalue(0)`.

`noorthog` suppresses orthogonal transformation of spline variables.

Reporting

`alleq` reports all equations used by `ml`. The models are fit using various constraints for parameters associated with the derivatives of the spline functions. These parameters are generally not of interest and thus are not shown be default. Also, an extra equation is used when fitting delayed-entry models; again, this is not shown by default.

`eform` reports the exponentiated coefficients. For models on the log cumulative-subdistribution hazard scale, `scale(hazard)`, this gives the subdistribution hazard ratios if the covariate is not time-dependent. Similarly, for models on the log cumulative-subdistribution odds scale, `scale(odds)`, this option will give odds ratios for non-time-dependent effects (see `scale()` option).

`level(#)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`; see [U] **23.5 Specifying the width of confidence intervals**.

Max options

`lininit` obtains initial values by fitting only the first spline basis function (i.e., a linear function of log survival time). This is useful when models fail to converge using the initial values obtained in the usual way. However, this option is seldom needed.

maximise_options `difficult`, `technique(`*algorithm_spec*`)`, `iterate(#)`, . These options are seldom used, but `difficult` may be useful if there are convergence problems when fitting more complicated models.

3.2 Equation Options

Model

`scale(`*scalename*`)` specifies the scale on which to model the cause-specific CIF.

`scale(hazard)` fits a model on the log-cumulative subdistribution hazards scale i.e. the scale of $\ln(-\ln(1-F_k(t)))$. If no time-dependent effects are specified, then the resulting model assumes proportionality.

`scale(odds)` fits a model on the log-cumulative odds scale i.e. the scale of $\log \frac{F_k(t)}{1-F_k(t)}$. If no time-dependent effects are specified, then the resulting model assumes proportionality of the odds ratios over time.

`df(#)` specifies the degrees of freedom for the restricted cubic spline function used for the baseline subdistribution hazard rate. Usually a value between 3 and 5 is sufficient and the choice of degrees of freedom has been shown not to be too sensitive to parameter estimates. Using `df(1)` is equivalent to fitting a Weibull model when using `scale(hazard)`. The internal knots are placed at the centiles of the distribution of the uncensored log times with boundary knots placed at the 0^{th} and 100^{th} centiles. An example is provided below for `df(5)`:

DF	Internal Knots	Centile Positions (Log-time)
5	4	20^{th} 40^{th} 60^{th} 80^{th}

`knots(`*numlist*`)` specifies knot locations for the baseline distribution function as opposed to the the default knot locations set by `df()`. The locations of the knots are placed on the log-time scale. Default knot positions are determined by the `df()` option.

`bknots(`*knotslist*`)` is a two-element list giving the boundary knots. By default, these are located at the minimum and maximum of the uncensored survival times for all cause-specific events on the log scale.

`tvc(`*varlist*`)` specifies the names of the variables that are time-dependent. Time-dependent effects are fit using restricted cubic splines. The degrees of freedom are specified using the `dftvc()` option.

`dftvc(`*df_list*`)` specifies the degrees of freedom for time-dependent effects. If the same degree of freedom is used for all time-dependent effects then the syntax is the same

as `df(#)`. With 1 degree of freedom, a linear effect of log-time is fit. If there is more than one time-dependent effect and different degrees of freedom are required for each time-dependent effect, then the following syntax can be used: `dftvc(x1:3 x2:2 1)`, where `x1` has 3 degrees of freedom, `x2` has 2 degrees of freedom, and any remaining time-dependent effects have 1 degree of freedom.

`knotstvc(`*numlist*`)` specifies the location of the internal knots for any time-dependent effects. If different knots are required for different time-dependent effects, then this option can be specified as follows: `knotstvc(x1 1 2 3 x2 1.5 3.5)`.

`cure` is specified when fitting cure models for a particular cause. It forces the cause-specific cumulative subdistribution hazard to be constant after the last knot. When the `df()` option is used together with the cure option, the internal knots are placed evenly according to centiles of the distribution of the uncensored log survival-times except one, which is placed at the 95^{th} centile and the final knot is placed outside of the last uncensored cause-specific log-survival time (110^{th} percentile by default). Alternative knot locations can be selected using the `knots()` option. Cure models can only be used when modelling on the log cumulative-subdistribution hazards scale (`scale(hazard)`).

`noconstant`; see [R] **estimation options**.

4 Postestimation

`stpm2cr` is an estimation command and shares most of the features of standard Stata estimation commands; see [U] **20 Estimation and postestimation commands**. The predictions available after fitting a modelling using `stpm2cr` are briefly described below.

4.1 Syntax

`predict` *newvarname* [*if*] [*in*] [`,` `at(`*varname* `#` [*varname* `#`]) `cause(`*numlist*`)` `chrdenominator(`*varname* `#` [*varname* `#` ...]) `chrnumerator(`*varname* `#` [*varname* `#` ...]) `ci cif cifdiff1(`*varname* `#` [*varname* `#` ...]) `cifdiff2(`*varname* `#` [*varname* `#` ...]) `cifratio csh cumodds cumsubhazard cured shrdenominator(`*varname* `#` [*varname* `#` ...]) `shrnumerator(`*varname* `#` [*varname* `#` ...]) `subdensity subhazard survivor timevar(`*varname*`)` `uncured xb zeros deviance dxb level(#)`]

Main

`at(`*varname* `#` [*varname* `#`]) requests that the covariates specified by *varname* be set to #. This is a useful way to obtain out-of-sample predictions. If `at()` is used together with `zeros`, then 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.

`cause(`*numlist*`)` specifies the causes on which to make the predictions for and are stored in *newvarname*_ c#. If `cause()` is not specified, then predictions are made for all causes included in the model and stored in *newvarname*_ c#.

`chrdenominator(`*varname* # [*varname* # ...]`)` and `shrdenominator(`*varname* # [*varname* # ...]`)` specifies the denominator of the cause-specific hazard ratio or subdistribution hazard ratio for a specific cause. By default, all covariates not specified using this option are set to zero. See the cautionary note in `chrnumerator()` and `shrnumerator` below. If # is set to missing (.), then the covariate has the values defined in the dataset.

`chrnumerator(`*varname* # [*varname* # ...]`)` `shrnumerator(`*varname* # [*varname* # ...]`)` specifies the numerator of the (time-dependent) cause-specific hazard ratio or subdistribution hazard ratio for a specific cause. By default, all covariates not specified using this option are set to zero. Setting the remaining values of the covariates to zero may not always be sensible, particularly on models other than those on the cumulative subdistribution hazard scale or when more than one variable has a time-dependent effect. If # is set to missing (.), then the covariate has the values defined in the dataset.

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

`cif` predicts the cause-specific cumulative incidence function.

`cifdiff1(`*varname* # [*varname* # ...]`)` and `cifdiff2(`*varname* # [*varname* # ...]`)` predict the difference in cause-specific cumulative incidence functions, with the first cause-specific cumulative incidence function defined by the covariate values listed for `cifdiff1()` and the second, by those listed for `cifdiff2()`. By default, covariates not specified using either option are set to zero. Setting the remaining values of the covariates to zero may not always be sensible. If # is set to missing (.), then *varname*has the values defined in the dataset.

Example: `cifdiff1(stage 1)` (without specifying `cifdiff2()`) computes the difference in predicted cause-specific cumulative incidence functions at `stage = 1` compared with `stage = 0` with all other covariates set to 0.

Example: `cifdiff1(stage 2) cifdiff1(stage 1)` computes the difference in predicted cause-specific cumulative incidence functions at `stage = 2` compared with `stage = 1`.

Example: `cifdiff1(stage 2 age 50) cifdiff1(stage 1 age 70)` computes the difference in predicted hazard functions at `stage = 2` and `age = 50` compared with `stage = 1` and `age = 70` with all other covariates set to 0.

`cifratio` predicts the relative contribution of failing from an event to the overall cumulative incidence function. For example, if the event of interest is in cancer, this

is the relative contribution of dying from cancer to the total mortality. `cifratio` must be used along with the `cause()` option in order to specify the cause-specific cumulative incidence function on the numerator of the ratio.

`csh` predicts the cause-specific hazard function.

`cumodds` predicts the cumulative odds-of-failure function.

`cumsubhazard` predicts the cumulative subdistribution hazard function.

`cured` predicts the cause-specific cure proportion after fitting a cure model.

`subdensity` predicts the sub-density function.

`subhazard` predicts the subdistribution hazard function.

`timevar(`*varname*`)` defines the variable used as time in the predictions. The default is `timevar(_t)`. The use of `timevar()` is useful for large datasets where, for plotting purposes, predictions are needed for only 200 observations, for example. Some caution should be taken when using this option because predictions may be made at whatever covariate values are in the first 200 rows of data. This can be avoided by using the `at()` option or the `zeros` option to define the covariate patterns for which you require the predictions.

`uncured` can be used after fitting a cure model for a specific cause. It can be used with the `survivor`, `subhazard`, and `cif` options to base predictions for the uncured group.

`xb` predicts the linear predictor, including the spline function.

`zeros` sets all covariates to zero (baseline prediction). For example, `predict cif, cause(1) cif zeros` calculates the baseline cause-specific cumulative incidence function for `cause = 1`.

Subsidiary

`dxb` calculates the derivative(s) of the linear predictor(s).

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

5 Examples

5.1 Northern European Cancer Registry Data (1975-94)

The methods outlined in this paper are illustrated through the use of Northern European cancer registry data, which has also been previously used to illustrate the use of `strs` for relative survival models (Dickman and Coviello 2015). We use a subset of this data which contains observations on 4,578 patients aged between 40 and 79 years old who were diagnosed with melanoma between 1975 and 1994. Survival time is measured

in months since diagnosis to death due to cancer or other causes. The covariates of interest are patient age at diagnosis and stage of cancer, which is categorised into localised or regional stage cancer at diagnosis. Follow-up time is restricted to 15 years from diagnosis.

5.2 Non-parametric estimates for the cause-specific cumulative incidence function

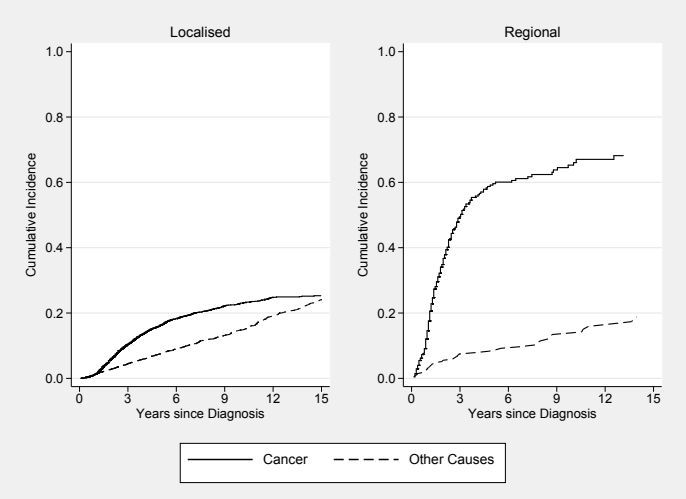

Figure 1: Predicted cause-specific cumulative incidence functions for death from cancer or death from other causes using the Aalen-Johansen method by stage at diagnosis for patients aged 40 to 80 years old.

Estimated cause-specific CIFs have been predicted using the `stcompet` command which implements the Aalen-Johansen method (Coviello and Boggess 2004). Figure 1 shows cause-specific CIFs estimated by stage at diagnosis for death from cancer and from other causes and shows that, those with a more distant stage cancer at diagnosis, have an increased risk of dying from cancer and lower risk of dying from other causes. The sum of the cancer-specific CIF and CIF for other causes give the overall, or all-cause probability of death.

5.3 Fine & Gray model

We initially fit direct regression models on the cause-specific CIF using the Fine & Gray approach which is, at present, the most commonly implemented method for modelling covariate effects on the cause-specific cumulative incidence function. Fine & Gray models are fitted with only stage at diagnosis as a covariate for each of the cause-specific CIFs.

A new indicator variable, `status2` was generated in order to overcome a small reporting error with the `stcrreg` command when using the `exit()` option in `stset` at the time of submission. When using the usual censoring indicator variable in `stset` for one cause before fitting a Fine & Gray model, because the competing events and censored events are no longer distinguished and those who die before the exit time are instead treated as censored, the number of actual competing events are under-reported. Although this has no direct consequence on the parameter estimates, the total number of overall failures that is reported for each cause-specific model is inconsistent. Therefore, we go on to fit Fine & Gray models using the new variable which is generated as shown below:

```
. stset surv_mm, failure(status == 1, 2) scale(12) id(id) exit(time 180)
  (output omitted)
. gen status2 = cond(_d==0,0,status)


. *Cancer
. stset surv_mm, failure(status2 == 1) scale(12) id(id) exit(time 180)
  (output omitted)
. stcrreg i.stage, compete(status2 == 2)
        failure _d:  status2 == 1
   analysis time _t:  surv_mm/12
  exit on or before:  time 180
                id:  id

Iteration 0:   log pseudolikelihood =  -7389.917
Iteration 1:   log pseudolikelihood = -7389.4747
Iteration 2:   log pseudolikelihood = -7389.4745

Competing-risks regression                    No. of obs       =      4,204
                                              No. of subjects  =      4,204
Failure event  : status2 == 1                 No. failed       =        937
Competing event: status2 == 2                 No. competing    =        583
                                              No. censored     =      2,684

                                              Wald chi2(1)     =     287.75
Log pseudolikelihood = -7389.4745             Prob > chi2      =     0.0000
                              (Std. Err. adjusted for 4,204 clusters in id)
------------------------------------------------------------------------------
             |               Robust
          _t |        SHR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |
   Regional  |   4.783974   .4414379    16.96   0.000     3.992499    5.732352
------------------------------------------------------------------------------

.
. *Other
. stset surv_mm, failure(status2 == 2) scale(12) id(id) exit(time 180)
  (output omitted)
. stcrreg i.stage, compete(status2 == 1)
        failure _d:  status2 == 2
   analysis time _t:  surv_mm/12
  exit on or before:  time 180
                id:  id

Iteration 0:   log pseudolikelihood = -4565.6556
Iteration 1:   log pseudolikelihood = -4556.6879
Iteration 2:   log pseudolikelihood = -4556.6578
```

```
Iteration 3:   log pseudolikelihood = -4556.6578
Competing-risks regression                      No. of obs       =      4,204
                                                No. of subjects  =      4,204
Failure event  : status2 == 2                   No. failed       =        583
Competing event: status2 == 1                   No. competing    =        937
                                                No. censored     =      2,684
                                                Wald chi2(1)     =       0.31
Log pseudolikelihood = -4556.6578               Prob > chi2      =     0.5790
                                (Std. Err. adjusted for 4,204 clusters in id)
------------------------------------------------------------------------------
             |               Robust
          _t |        SHR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       stage |
   Regional  |   .9080851   .1577827    -0.55   0.579     .6459927    1.276514
------------------------------------------------------------------------------
```

The SHR for cancer gives the association between stage at diagnosis and the cancer-specific CIF. A SHR of 4.78 indicates that, those with a more severe stage at diagnosis is associated with an increased risk of dying from cancer. However, it is important to note that, due to the awkward definition in the risk-set, it is difficult to make inferences on quantitative effects. Although non-significant, the SHR from the Fine & Gray model for other causes shows that, those with a more severe stage at diagnosis is associated with a decreased risk of dying from other causes. This is explained by the fact that patients at an earlier stage at diagnosis are healthier and are more likely to live longer and die of other causes before their cancer. Whereas, on the other hand, patients at a later stage are unlikely to live as long to have the chance of dying from other causes.

After fitting each cause-specific Fine & Gray model, `stcurve` can be used to predict and store the cause-specific CIFs. These are stored and plotted later in Figure 3.

5.4 Log-cumulative subdistribution hazard models

Using the full-likelihood in Equation 11, direct flexible parametric regression models for the cause-specific CIF can be fitted. Rather than fitting a model to each cause-specific CIF separately, this approach allows to instead model all cause-specific CIFs simultaneously. This is shown below with the assumption of proportionality for all causes:

```
. stset surv_mm, failure(status==1, 2) scale(12) id(id) noshow exit(time 180)
  (output omitted)
. stpm2cr [cancer: stage2, scale(hazard) df(5) ] ///
>                 [other: stage2, scale(hazard) df(5) ] ///
>                 , events(status) cause(1 2) cens(0) eform nolog
  (output omitted)
Log likelihood = -4901.0253                     Number of obs     =      4,204
------------------------------------------------------------------------------
             |      exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
cancer       |
      stage2 |   4.673522   .3973545    18.14   0.000     3.956153    5.520973
------------------------------------------------------------------------------
```

_rcs_c1_1	2.371601	.0642335	31.88	0.000	2.248989	2.500897
_rcs_c1_2	1.40679	.0445023	10.79	0.000	1.322216	1.496774
_rcs_c1_3	1.061522	.0237518	2.67	0.008	1.015975	1.109111
_rcs_c1_4	.9889806	.0103402	-1.06	0.289	.9689204	1.009456
_rcs_c1_5	1.002836	.005948	0.48	0.633	.9912455	1.014562
_cons	.1390518	.0053603	-51.18	0.000	.1289329	.1499648
other						
stage2	.6867003	.115223	-2.24	0.025	.4942449	.9540964
_rcs_c2_1	2.564841	.0949475	25.44	0.000	2.385338	2.757852
_rcs_c2_2	1.058082	.0298144	2.00	0.045	1.001231	1.118161
_rcs_c2_3	.9541731	.0196412	-2.28	0.023	.9164434	.9934562
_rcs_c2_4	.9843678	.0125716	-1.23	0.217	.9600337	1.009319
_rcs_c2_5	.9917352	.0082375	-1.00	0.318	.9757208	1.008012
_cons	.0800586	.0040859	-49.47	0.000	.0724379	.088481

An equation is specified for each cause within the square brackets along with their respective options. These are similar to those used for `stpm2` where `df(5)` implies 4 internal knots at default locations. The estimated subdistribution hazard ratios are displayed for each cause and their 95% confidence intervals. The advantage of using the parametric approach is that it is easy to obtain useful predictions to aid interpretation. The following code obtains the cause-specific CIFs, subdistribution hazard functions for each cause and the cause-specific hazard functions. Confidence intervals are obtained by using the `ci` option.

```
. range temptime 0 15 1000
. predict cif1, cif at(stage1 1 stage2 0) timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict cif2, cif at(stage1 0 stage2 1) timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict sdh1, subhazard at(stage1 1 stage2 0) timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict sdh2, subhazard at(stage1 0 stage2 1) timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict csh1, csh at(stage1 1 stage2 0) timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict csh2, csh at(stage1 0 stage2 1) timevar(temptime)
Calculating predictions for the following causes: 1 2
```

The top row in Figure 2 plots the predicted subdistribution hazard function for each cause and the bottom illustrates the predicted cause-specific hazard function by stage at diagnosis. The subdistribution hazard gives the association on the effect of stage at diagnosis on risk and the cause-specific hazard is the association on the effect of stage at diagnosis on the hazard rate. Figure 3 compares the cause-specific CIFs obtained from the Fine & Gray models for each cause to those obtained from the log-cumulative proportional subdistribution hazards model and shows sensible agreement between the two (refer to Mozumder et al. (2016 (submitted) for more details on the disagreement in the cause-specific CIF for death from other causes). In Figure 4 the Aalen-Johansen estimates are compared to the cause-specific CIFs obtained from the log-cumulative proportional subdistribution hazards model. The estimates are reasonably similar, however, a better fit can be achieved by relaxing the assumption of proportionality through the

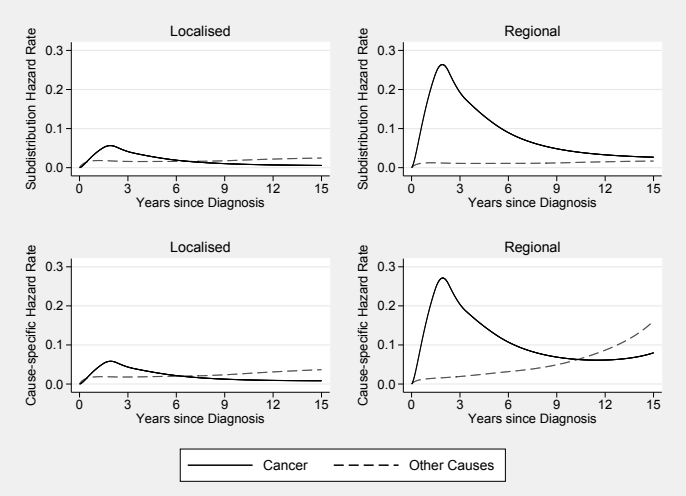

Figure 2: Subdistribution hazards predicted for each cause and cause-specific hazard predictions by stage at diagnosis for patients aged 40 to 80 years old from a log-cumulative proportional subdistribution hazards model for melanoma data.

inclusion of time-dependent effects using restricted cubic splines.

5.5 Time-dependent effects

The inclusion of time-dependent effects can be easily incorporated by specifying the `dftvc()` and `tvc()` equation specific options as shown in the following code.

```
. stpm2cr [cancer: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
>                 [other: stage2, scale(hazard) df(5) tvc(stage2) dftvc(3)] ///
>                 , events(status) cause(1 2) cens(0) eform nolog
  (output omitted)
Log likelihood = -4877.5917                     Number of obs     =      4,204
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cancer						
stage2	5.225629	.4543429	19.02	0.000	4.406875	6.196499
_rcs_c1_1	2.570244	.089602	27.08	0.000	2.400493	2.752
_rcs_c1_2	1.440213	.0605618	8.68	0.000	1.326274	1.56394
_rcs_c1_3	1.076737	.0280174	2.84	0.004	1.023201	1.133074
_rcs_c1_4	.9907888	.0106845	-0.86	0.391	.9700674	1.011953
_rcs_c1_5	.9997375	.0058897	-0.04	0.964	.9882603	1.011348
_rcs_stage2_c1_1	.7353858	.0413674	-5.46	0.000	.658617	.8211029
_rcs_stage2_c1_2	.9750149	.0568817	-0.43	0.664	.8696665	1.093125
_rcs_stage2_c1_3	.9458115	.0303569	-1.74	0.083	.8881458	1.007221
_cons	.1328929	.0053238	-50.38	0.000	.1228576	.143748
other						
stage2	1.18831	.2267027	0.90	0.366	.8175976	1.727109

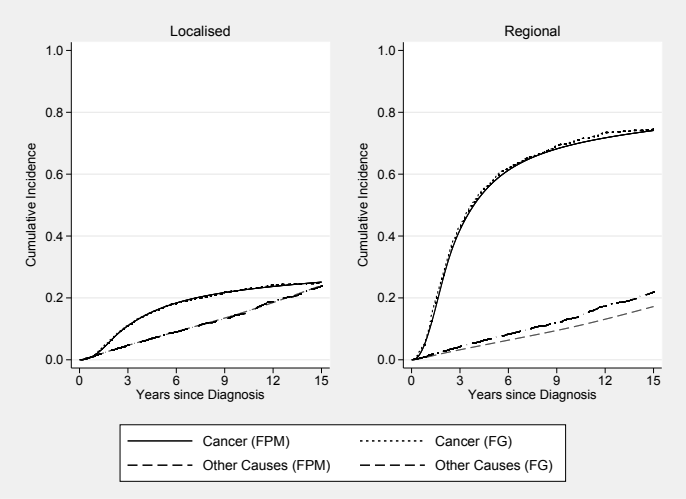

Figure 3: A comparison of cause-specific cumulative incidence functions for death from cancer or death from other causes predicted simulataneously from a log-cumulative subdistribution hazards model and from seperate Fine & Gray models for each cause by stage at diagnosis for patients aged 40 to 80 years old.

_rcs_c2_1	2.658485	.1059802	24.53	0.000	2.458675	2.874533
_rcs_c2_2	1.062388	.0328778	1.96	0.051	.999864	1.128822
_rcs_c2_3	.9584928	.0206057	-1.97	0.049	.9189454	.9997422
_rcs_c2_4	.9841378	.0124521	-1.26	0.206	.9600322	1.008849
_rcs_c2_5	.9926364	.0081918	-0.90	0.370	.9767099	1.008823
_rcs_stage2_c2_1	.68066	.0697333	-3.75	0.000	.5568331	.8320231
_rcs_stage2_c2_2	1.007956	.0739275	0.11	0.914	.8729933	1.163783
_rcs_stage2_c2_3	.9515855	.0501094	-0.94	0.346	.8582712	1.055045
_cons	.0775996	.0040571	-48.89	0.000	.0700417	.0859732

The `tvc(stage2)` and `dftvc(3)` options states that the `stage2` variable is to be time-dependent using restricted cubic splines with 2 internal knots (i.e. 3 degrees of freedom). Overall, there are 10 parameters being estimated for each cause in the model. For example, for cancer, there are 5 derived variables for the baseline log-cumulative subdistribution hazard (`_rcs_c1_1-_rcs_c1_5`) and 3 derived splines for the time-dependent effect `stage2` (`_rcs_stage2_c1_1-_rcs_stage2_c1_3`).

In a time-dependent model, parameter estimates become more complex and are not very useful when interpreted on their own. Instead, it is better to obtain predictions between groups for specific covariate patterns as relative and/or absolute differences over time by using `predict`. Note here that, to generate the same predictions, the coding is the same:

```
. range temptime 0 15 1000
. predict cif_tvc1, cif at(stage1 1 stage2 0) ci timevar(temptime)
Calculating predictions for the following causes: 1 2
```

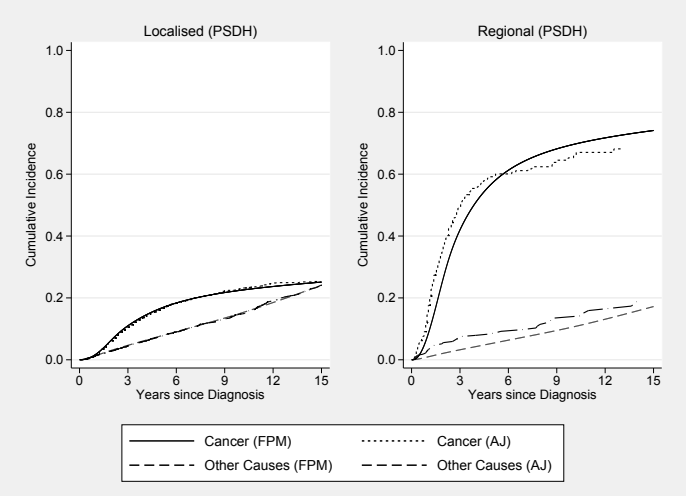

Figure 4: A comparison of cause-specific cumulative incidence functions for death from cancer or death from other causes predicted simulataneously from a log-cumulative subdistribution hazards model assuming proportionality and using the Aalen-Johansen empirical estimates for each cause by stage at diagnosis for patients aged 40 to 80 years old.

```
. predict cif_tvc2, cif at(stage1 0 stage2 1) ci timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict cifdiff, cifdiff1(stage1 0 stage2 1) cifdiff2(stage1 1 stage2 0) ci timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict shr, shrn(stage1 0 stage2 1) shrd(stage1 1 stage2 0) ci timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict chr, chrn(stage1 0 stage2 1) chrd(stage1 1 stage2 0) ci timevar(temptime)
Calculating predictions for the following causes: 1 2
```

Figure 5 now shows a better fit of the model estimated cause-specific CIFs, particularly with regional stage patients, in comparison to the non-parametric Aalen-Johansen estimates with very good agreement.

We can obtain absolute differences with 95% confidence intervals between the regional and localised stage groups over time for each cause-specific CIF. Differences are calculated by using the `cifdiff1()` and `cifdiff2()` options. The obtained predictions are illustrated in Figure 6, which show us that, those with a more severe stage of cancer at diagnosis, are more likely to die from cancer. The difference is smaller for other causes for the first 6 years since diagnosis and in the later years, the cause-specific CIF for other causes is larger for localised stage patients.

Time-dependent subdistribution and cause-specific hazard ratios are obtained using the options, `shrnumerator()` and `shrdenominator()`, and `chrnumerator()` and `chrdenominator()` respectively. Using these options, we can obtain ratios for any

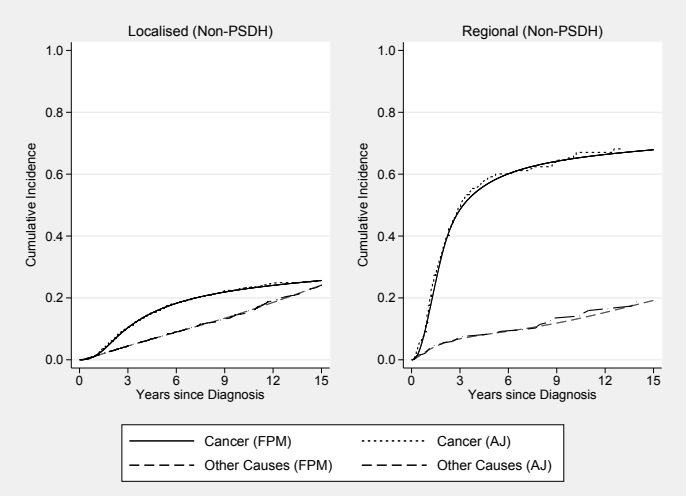

Figure 5: A comparison of cause-specific cumulative incidence functions for death from cancer or death from other causes predicted simulataneously from a log-cumulative non-proportional subdistribution hazards model and using the Aalen-Johansen empirical estimates for each cause by stage at diagnosis for patients aged 40 to 80 years old.

two covariate patterns. Figure 7 shows the time-dependent subdistribution and cause-specific hazard ratios and compares regional stage patients to localised stage patients at diagnosis. At the start of follow up, for both cancer-specific hazard ratios, regional stage patients have a mortality rate that is 17 times the mortality rate of localised stage patients and decreases over follow up time. The mortality rate due to other causes on both scales for regional stage patients at the start of follow up time is approximately 4.5 times that of localised stage patients. Beyond 2 years since diagnosis, the subdistribution hazard rate due to other causes for regional stage patients is lower than the localised stage patients since the ratio is less than 1 which is expected since those at a later stage will die earlier due to the cancer before they have the chance to die of other causes. The cause-specific hazard ratios give us the association of stage at diagnosis on the rate. The CSHR show a different effect on death due to other causes because patients at a later stage tend to be more sick and, in general, are at a higher risk of dying. This translates to a positive association between more distant stage patients and the mortality rate for other causes.

5.6 Cure model

Cure models for any causes can be fit by adding the equation option `cure`, however, it is highly recommended that this is done for one cause, which is usually the event of interest. Predictions can be made after fitting a cure model with `predict` using the `cured` and `uncured` options. Specifying the `cured` option will calculate the cure proportion for the

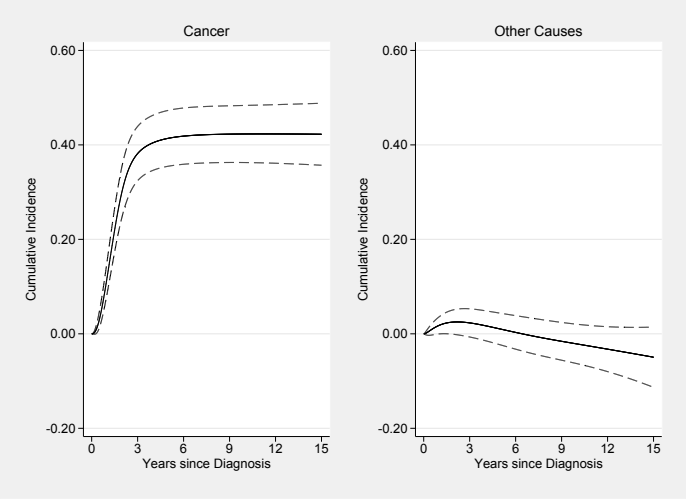

Figure 6: Predicted absolute differences (Regional - Localised) in cause-specific cumulative incidence functions with 95% confidence intervals from a log-cumulative non-proportional subdistribution hazards model.

cause that cure was specified for and a variable with the suffix `_btd` that partitions those that are still alive into two groups; patients bound to die from cancer and not bound to die from cancer. The code for fitting a cure model and predictions are shown below:

```
. stpm2cr [cancer: , scale(hazard) df(5) cure] ///
>                 [other: , scale(hazard) df(5)] ///
>                 , events(status) cause(1 2) cens(0) eform mlmethod(lf2) nolog
  (output omitted)
Log likelihood = -1742.7601                     Number of obs     =      1,692
```

	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cancer						
_rcs_c1_1	2.168448	.0851865	19.70	0.000	2.007752	2.342007
_rcs_c1_2	.9134977	.0245224	-3.37	0.001	.8666772	.9628475
_rcs_c1_3	.9989706	.0182824	-0.06	0.955	.9637729	1.035454
_rcs_c1_4	.9775022	.0134488	-1.65	0.098	.9514954	1.00422
_rcs_c1_5	1	(omitted)				
_cons	.348136	.0181445	-20.25	0.000	.3143294	.3855784
other						
_rcs_c2_1	2.645041	.3083898	8.34	0.000	2.104696	3.324111
_rcs_c2_2	.9981758	.0919501	-0.02	0.984	.8332895	1.195689
_rcs_c2_3	.9368575	.0517331	-1.18	0.238	.8407566	1.043943
_rcs_c2_4	1.013603	.037129	0.37	0.712	.9433826	1.089051
_rcs_c2_5	.9643029	.0211338	-1.66	0.097	.9237584	1.006627
_cons	.0220712	.0032665	-25.77	0.000	.0165139	.0294985

```
. range temptime 0 15 1000
```

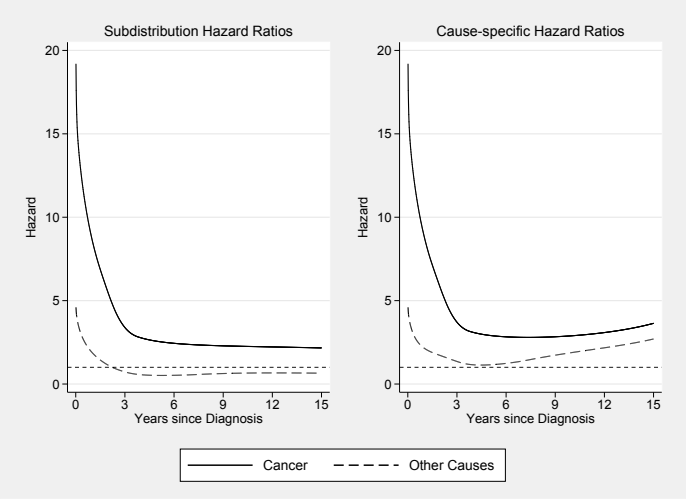

Figure 7: Predicted subdistribution and cause-specific hazard ratios for each cause from a log-cumulative non-proportional subdistribution hazards model. Ratios compare regional stage to localised stage patients at diagnosis. Dotted line is a reference line when the rate is equal to 1 i.e. no difference.

```
. predict cif, cif timevar(temptime)
Calculating predictions for the following causes: 1 2
. predict cure, cured timevar(temptime)
Calculating predictions for the following causes: 1 2
. gen cif_tot = cif_c1 + cif_c2
```

In Section 2.8, we showed that, to fit cure models, the last knot was constrained to be zero to force a plateau. This is shown in the output above where the parameter for `_rcs_c1_5` is equal to one. Analysis is restricted to localised stage patients aged 40 to 54 years old where cure is found to be reasonable. To check this, the plot to the left in Figure 8 compares the estimated cancer-specific CIF from the model with the Aalen-Johansen estimate and shows extremely good agreement with the cure proportion estimated at approximately 30% after 12 years since diagnosis where the cancer-specific CIF plateaus. On the right hand side of Figure 8, the cause-specific CIFs are stacked and the dashed line is the partitioning of alive patients that are bound to or not bound to die into two groups. This estimate is provided as part of the `cured` option with the suffix `_btd`. Eloranta et al. (2014) introduces this quantity to aid better risk communication and is calculated as follows:

$$P_{alive,can}(t) = \pi_c - F_1(t) \tag{20}$$

$$P_{alive,oth}(t) = 1 - F_2(t) - \cdots - F_K(t) - \pi_c \tag{21}$$

where π_c is the proportion of those bound-to-die from cancer on which cure is assumed. For $k = 1$, $P_{alive,can}(t)$ represents patients who will ultimately die from their cancer, and $P_{alive,oth}(t)$ give those who will die from competing causes where $k = 2, \cdots, K$. In our example, from the stacked probabilities in Figure 8, at 6 years after diagnosis, approximately 25% have died and 6% are alive and bound to die from cancer, and 69% are alive and not bound to die from cancer.

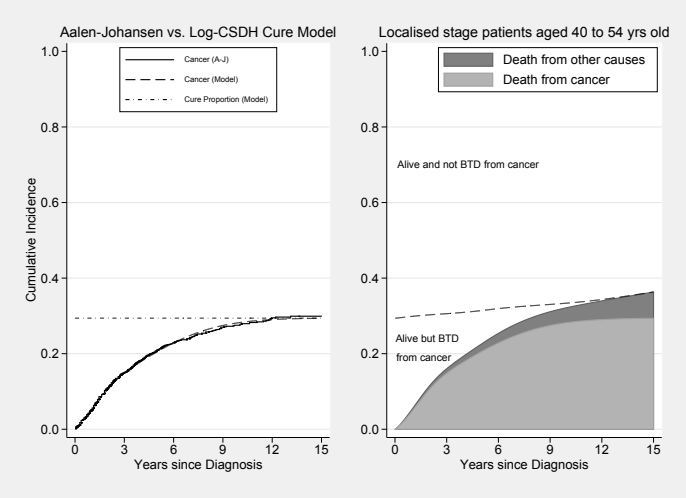

Figure 8: Left: Comparison of predicted cancer-specific CIFs obtained from log-cumulative subdistribution hazards cure model and using the Aalen-Johansen method for localised stage patients aged 40 to 54 years old. Right: Stacked cause-specific CIFs obtained from a log-cumualtive subdistribution hazards cure model. Dashed-line partitions patients who are still alive into those who are bound to die (BTD) from cancer and not BTD from cancer.

5.7 Conclusions

Competing risks models are being more widely applied in research and fitting regression models on the subdistribution hazard scale is encouraged to make inferences on prognosis and understand the association of a covariate on risk. Analysis from within the flexible parametric modelling framework using the direct likelihood approach for the cause-specific CIF has several advantages. This includes computational time gains as numerical integration is not required to model the baseline log-cumulative subdistribution hazard function and all causes are modelled simultaneously so there is no need to fit separate models for each cause. This is implemented in the new `stpm2cr` command, which is an adaptation of the `stpm2` command. Other useful predictions can be obtained by using `predict` after fitting a model using `stpm2cr`. This complements flexible parametric regression models for competing risks on the cause-specific hazard scale and allows researchers to gain a more complete understanding on the impact of the

event of interest on outcome. There is scope for further improvement of these models as convergence can be difficult if they are misspecified and mortality in a covariate group is high which may cause the sum of all probabilities to exceed one. Therefore, future work may involve implementing an appropriate constraint on the models to avoid issues in convergence.

6 References

Andersen, P. K., R. B. Geskus, T. de Witte, and H. Putter. 2012. Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* 41(3): 861–870.

Andersson, T. M., P. W. Dickman, S. Eloranta, and P. C. Lambert. 2011. Estimating and modelling cure in population-based cancer studies within the framework of flexible parametric survival models. *BMC medical research methodology* 11(1): 1.

Beyersmann, J., and M. Schumacher. 2007. Misspecified regression model for the subdistribution hazard of a competing risk. *Statistics in medicine* 26(7): 1649.

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

Dickman, P. W., and E. Coviello. 2015. Estimating and modelling relative survival. *The Stata Journal* 15(1): 186–215. http://www.stata-journal.com/article.html?article=st0376.

Durrleman, S., and R. Simon. 1989. Flexible regression models with cubic splines. *Statistics in medicine* 8(5): 551–561.

Eloranta, S., P. C. Lambert, T. M.-L. Andersson, M. Björkholm, and P. W. Dickman. 2014. The application of cure models in the presence of competing risks: a tool for improved risk communication in population-based cancer patient survival. *Epidemiology* 25(5): 742–748.

Fine, J. P., and R. J. Gray. 1999. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* 94(446): 496–509.

Gerds, T. A., T. H. Scheike, and P. K. Andersen. 2012. Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in medicine* 31(29): 3921–3930.

Gray, R. J. 1988. A class of K-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of statistics* 1141–1154.

Hinchliffe, S. R., and P. C. Lambert. 2013. Flexible parametric modelling of cause-specific hazards to estimate cumulative incidence functions. *BMC medical research methodology* 13(1): 1.

Jeong, J.-H., and J. Fine. 2006. Direct parametric inference for the cumulative incidence function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 55(2): 187–200.

Koller, M. T., H. Raatz, E. W. Steyerberg, and M. Wolbers. 2012. Competing risks and the clinical community: irrelevance or ignorance? *Statistics in medicine* 31(11-12): 1089–1097.

Lambert, P., W. S. R., and M. Crowther. 2016 (submitted). Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in Medicine* .

Lambert, P. C., L. Holmberg, F. Sandin, F. Bray, K. M. Linklater, A. Purushotham, D. Robinson, and H. Møller. 2011. Quantifying differences in breast cancer survival between England and Norway. *Cancer Epidemiology* 35: 526–533. http://dx.doi.org/10.1016/j.canep.2011.04.003.

Lambert, P. C., and P. Royston. 2009. Further development of flexible parametric models for survival analysis. *The Stata Journal* 9: 265–290.

Lau, B., S. R. Cole, and S. J. Gange. 2009. Competing risk regression models for epidemiologic data. *American journal of epidemiology* kwp107.

Mozumder, S. I., M. Rutherford, and P. Lambert. 2016 (submitted). Direct likelihood inference on the cause-specific cumulative incidence function: a flexible parametric regression modelling approach. *Statistics in Medicine* .

Noordzij, M., K. Leffondré, K. J. van Stralen, C. Zoccali, F. W. Dekker, and K. J. Jager. 2013. When do we need competing risks methods for survival analysis in nephrology? *Nephrology Dialysis Transplantation* 28(11): 2670–2677.

Putter, H., M. Fiocco, and R. Geskus. 2007. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in medicine* 26(11): 2389–2430.

Royston, P., and M. K. Parmar. 2002. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine* 21(15): 2175–2197.

Sapir-Pichhadze, R., M. Pintilie, K. Tinckam, A. Laupacis, A. Logan, J. Beyene, and S. Kim. 2016. Survival analysis in the presence of competing risks: the example of wait-listed kidney transplant candidates. *American Journal of Transplantation* .

Wolbers, M., M. T. Koller, V. S. Stel, B. Schaer, K. J. Jager, K. Leffondré, and G. Heinze. 2014. Competing risks analyses: objectives and approaches. *European heart journal* ehu131.

About the authors

Sarwar Islam Mozumder is a PhD student at the University of Leicester, UK. His PhD focusses on further development of flexible parametric modelling methods in competing risks, risk communication of cancer survival statistics and maximising the use of more detailed population-based colorectal data.

Paul Lambert is a Professor of Biostatistics at the University of Leicester, UK. He also works part-time at the Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. His main interest is in the development and application of methods in population-based cancer research.

Mark Rutherford is a Lecturer of Biostatistics at the University of Leicester, UK. He has a keen interest in applying survival methods in Stata, particularly for population-based cancer data.