The Stata Journal

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

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

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

http://www.stata-journal.com

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

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

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

U.S. and Canada		Elsewhere	
1-year subscription	$ 79	1-year subscription	$115
2-year subscription	$155	2-year subscription	$225
3-year subscription	$225	3-year subscription	$329
3-year subscription (electronic only)	$210	3-year subscription (electronic only)	$210
1-year student subscription	$ 48	1-year student subscription	$ 79
1-year university library subscription	$ 99	1-year university library subscription	$135
2-year university library subscription	$195	2-year university library subscription	$265
3-year university library subscription	$289	3-year university library subscription	$395
1-year institutional subscription	$225	1-year institutional subscription	$259
2-year institutional subscription	$445	2-year institutional subscription	$510
3-year institutional subscription	$650	3-year institutional subscription	$750

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

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

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

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

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

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

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Stata, Stata Press, Mata, Mata, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 4, pp. 623–638

Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models

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

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

Abstract. When the mortality among a cancer patient group returns to the same level as in the general population, that is, when the patients no longer experience excess mortality, the patients still alive are considered "statistically cured". Cure models can be used to estimate the cure proportion as well as the survival function of the "uncured". One limitation of parametric cure models is that the functional form of the survival of the uncured has to be specified. It can sometimes be hard to find a survival function flexible enough to fit the observed data, for example, when there is high excess hazard within a few months from diagnosis, which is common among older age groups. This has led to the exclusion of older age groups in population-based cancer studies using cure models. Here we use flexible parametric survival models that incorporate cure as a special case to estimate the cure proportion and the survival of the uncured. Flexible parametric survival models use splines to model the underlying hazard function; therefore, no parametric distribution has to be specified. We have updated the `stpm2` command for flexible parametric models to enable cure modeling.

Keywords: st0165_1, stpm2, stpm2 postestimation, cure models, flexible parametric survival model, relative survival, survival analysis

1 Introduction

Patient survival, the time from diagnosis to death, is the most important single measure of cancer patient care (the diagnosis and treatment of cancer). Cancer patient survival

is often measured using five-year relative survival, an estimate of net survival, that is, the proportion of patients who would still be alive five years after diagnosis if the cancer (and anything directly or indirectly related to the cancer) was the only possible cause of death (Dickman and Adami 2006). Relative survival is estimated as the observed survival divided by the expected survival and can be interpreted as net survival under the assumption that the cancer patients would have the same expected survival as the general population if they had not had cancer. Because cancer patient survival has improved for many cancer types, and many patients are cured of their disease, it is also important to find the proportion of patients that are cured of their cancer.

For most cancers, the relative survival will reach a plateau some years after diagnosis, indicating that the mortality among the patients still alive is the same as expected in the general population. This point is called the cure point, and the patients still alive are considered "statistically cured". De Angelis et al. (1997), Verdecchia et al. (1998), Yu et al. (2004), and Lambert et al. (2007b) have proposed cure models for population-based cancer studies that can be used to estimate the proportion of cancer patients statistically cured.

There are two user-written Stata commands for cure modeling within a relative survival setting: `strsmix` and `strsnmix`. These commands can be used to fit mixture and nonmixture cure models, respectively. These models are parametric cure models. The models give estimates of the cure proportion as well as the survival of those "uncured". These measures are of interest to patients, clinicians, and policy makers and can give valuable insights into temporal trends in cancer patient survival. One limitation of parametric cure models is that the functional form of the survival of the uncured has to be specified. It can sometimes be difficult to fit survival functions flexible enough to capture high excess hazard within a few months from diagnosis, which is common among older age groups. This has led to the exclusion of older age groups in population-based cancer studies using cure models (Lambert et al. 2007a). In our experience, the current models can also give biased estimates, or fail to converge, when the cure proportion is high (for example, 80% and above).

This problem can be avoided by using flexible parametric survival models to estimate the cure proportion and the survival of the uncured in a population-based setting. Flexible parametric survival models were first introduced by Royston (2001) and Royston and Parmar (2002) and extended to relative survival by Nelson et al. (2007) and Lambert and Royston (2009). The models use splines to fit the underlying distribution and can therefore more easily capture the shape. The `stpm2` command can be used to fit flexible parametric models for both cause-specific and relative survival, and it has now been updated to incorporate the flexible parametric cure model for relative survival as described by Andersson et al. (2011).

2 Model specifications

2.1 Relative survival and excess mortality

The method of choice for studying cancer patient survival in a population-based setting is relative survival, $R(t)$ (Dickman and Adami 2006). Relative survival is the observed (all-cause) survival, $S(t)$, among the cancer patients divided by the expected survival, $S^*(t)$, in a hypothetical group in the general population that is comparable to the cancer patients with respect to age, sex, calendar year, and possible other covariates. An advantage of relative survival is that it does not rely on classification of cause of death, which is known to be poorly reported (Begg and Schrag 2002). In the relative survival model, the overall survival can be written as

$$S(t) = S^*(t)R(t)$$

The hazard analogue of relative survival is the excess hazard rate. The overall hazard, $h(t)$, among the patients is the sum of two components: the expected hazard, $h^*(t)$, and the excess hazard, $\lambda(t)$, associated with a diagnosis of the cancer.

$$h(t) = h^*(t) + \lambda(t) \tag{1}$$

Both $S^*(t)$ and $h^*(t)$ are assumed known and are usually obtained from routine data sources (for example, national or regional life tables).

2.2 Flexible parametric survival model

The flexible parametric survival model (Nelson et al. 2007; Lambert and Royston 2009) for relative survival is fit on the log cumulative-excess-hazard scale by using restricted cubic splines to estimate the baseline cumulative excess hazard. By integrating (1), we get

$$H(t) = H^*(t) + \Lambda(t)$$

where $H(t)$ is the overall cumulative hazard, $H^*(t)$ is the expected cumulative hazard, and $\Lambda(t)$ is the cumulative excess hazard. We model the cumulative excess hazard on the log scale by using restricted cubic splines:

$$\ln\{\Lambda(t)\} = \ln\{-\ln R(t)\} = \gamma_{00} + \gamma_{01}v_1(x) + \gamma_{02}v_2(x) + \cdots + \gamma_{0K-1}v_{K-1}(x) \tag{2}$$

where $x = \ln(t)$, K is the number of knots, $v_1(x) = x$, and for $j = 2, \ldots, K-1$, the jth basis function is defined as

$$v_j(x) = (x - k_j)_+^3 - \lambda_j(x - k_{\min})_+^3 - (1 - \lambda_j)(x - k_{\max})_+^3$$

where $u_+ = u$ if $u > 0$ and $u_+ = 0$ if $u \leq 0$, $k_{\min}$ is the position of the first knot, $k_{\max}$ the position of the last knot, and $\lambda_j = (k_{\max} - k_j)/(k_{\max} - k_{\min})$.

Introducing covariates, $\mathbf{z}$, into (2) gives

$$\begin{aligned}\ln\{\Lambda(t;\mathbf{z})\} &= \ln\{-\ln R(t;\mathbf{z})\} \\ &= \gamma_{00} + \gamma_{01}v_1(x) + \gamma_{02}v_2(x) + \cdots + \gamma_{0K-1}v_{K-1}(x) + \boldsymbol{\beta}^T\mathbf{z}\end{aligned}$$

This is a proportional excess-hazards model. Nonproportional excess-hazards models, that is, models with time-dependent covariate effects, can be modeled by including interactions between covariates and splines for time

$$\ln\{\Lambda(t;\mathbf{z})\} = s(x;\boldsymbol{\gamma_0}) + \boldsymbol{\beta}^T\mathbf{z} + \sum_{i=1}^{D} s(x;\boldsymbol{\gamma_i})z_i$$

where $s(x;\boldsymbol{\gamma_0})$ is the spline function for the baseline log cumulative excess hazard as expressed in (2), D is the number of time-dependent covariate effects, and $s(x;\boldsymbol{\gamma_i})$ is the spline function for the ith time-dependent effect.

2.3 Parametric cure models for relative survival

For most cancers, the mortality in the patient group will, after some years from diagnosis, return to the same level as in the general population; that is, $\lambda(t)$ in (1) is equal to 0 after some point. This point is called the cure point, and the patients still alive are considered statistically cured. This is a population definition of "cured" and does not necessarily imply that all patients are medically cured. Statistical cure is a useful method of measuring long-time survival in population-based cancer studies.

One of the most often used cure models in population-based cancer studies is the mixture cure model (Verdecchia et al. 1998; De Angelis et al. 1997; Lambert et al. 2007b). When one incorporates relative survival, the overall survival function from the mixture cure model can be written as

$$S(t) = S^*(t)\{\pi + (1-\pi)S_u(t)\}$$

It assumes that a proportion, π, of the patients will be cured (will not experience excess mortality), while the remainder, $1-\pi$, are uncured. $S_u(t)$ is the cancer-specific survival function for the uncured and is estimated by the model along with the cure proportion. A parametric distribution for $S_u(t)$ has to be chosen, and a Weibull distribution is often used (De Angelis et al. 1997; Verdecchia et al. 1998; Lambert et al. 2007a,b).

Another parametric cure model used in population-based cancer studies is the nonmixture cure model (Lambert et al. 2007b), which fits an asymptote for the survival function at the cure proportion. The survival function for the nonmixture model can be written as

$$S(t) = S^*(t)\pi^{F_Z(t)}$$

where $F_Z(t)$ is a cumulative distribution function, generally chosen to be $1 - S_Z(t)$, where $S_Z(t)$ is a standard parametric survival function. As for the mixture model, a Weibull distribution is often used. Thus the relative survival function has an asymptote at the cure fraction, π. The nonmixture model can be written as a mixture model:

$$S(t) = S^*(t)\left\{\pi + (1-\pi)\left(\frac{\pi^{F_Z(t)} - \pi}{1-\pi}\right)\right\}$$

This enables estimation of both the cure proportion and the survival of the uncured. In modeling, both the cure proportion and the parameters in $S_u(t)$ or $F_Z(t)$ can be allowed to vary by covariates.

2.4 Flexible parametric cure models

When cure is reached, the excess-hazard rate is 0, and the cumulative excess hazard will be constant after this time. By forcing the log cumulative excess hazard in the flexible parametric survival model not only to be linear but also to have zero slope after the last knot, we can estimate the cure proportion. We do this by calculating the spline variables backward (that is, treating the knots in reverse order) and then by restricting the parameter for the linear spline variable to 0 ($\gamma_{01} = 0$). The spline basis functions, $v_j(x)$, are then defined as

$$v_j(x) = (k_{K-j+1} - x)_+^3 - \lambda_j(k_{\max} - x)_+^3 - (1 - \lambda_j)(k_{\min} - x)_+^3$$

for $j = 2, \ldots, K-1$, and $\lambda_j = (k_{K-j+1} - k_{\min})/(k_{\max} - k_{\min})$. The relative survival function from the flexible parametric survival model, with splines calculated backward and with restriction for the linear spline variable, is defined as

$$R(t) = \exp\left[-\exp\left\{\gamma_{00} + \gamma_{02}v_2(x) + \cdots + \gamma_{0K-1}v_{K-1}(x)\right\}\right]$$

which can be written as

$$R(t) = \pi^{\exp\{\gamma_{02}v_2(x) + \cdots + \gamma_{0K-1}v_{K-1}(x)\}}$$

where $\pi = \exp\{-\exp(\gamma_{00})\}$. This is a special case of a nonmixture cure model: the cure proportion is $\pi = \exp\{-\exp(\gamma_{00})\}$; the distribution function is $F_Z(t) = \exp\{\gamma_{02}v_2(x) + \cdots + \gamma_{0K-1}v_{K-1}(x)\}$.

When one includes covariates,

$$R(t;\mathbf{z}) = \exp\left[-\exp(\gamma_{00} + \boldsymbol{\beta}^T\mathbf{z})\exp\left\{\gamma_{02}v_2(x) + \cdots + \gamma_{0K-1}v_{K-1}(x) + \sum_{i=1}^{D} s(x;\boldsymbol{\gamma_i})z_i\right\}\right] \quad (3)$$

the constant parameters, γ_{00} and $\boldsymbol{\beta}$, are used to model the cure proportion, and the time-dependent parameters are used to model the distribution function $F_Z(t)$. The constraint of a zero effect for the linear spline term has to be incorporated for each spline function that we model.

All spline variables take the value 0 from the point of the last knot, which means that in (3), the constant parameter, γ_{00}, is the log cumulative excess hazard at and beyond the last knot for the reference group and can therefore be used to predict cure. It is usually preferred to orthogonalize the spline variables, which results in them not being 0 from the point of the last knot. Because of this, cure cannot be predicted by a direct transformation of the constant parameters. Therefore, we have chosen to center

the orthogonalized spline variables around the value they take at the last knot, which enables direct predictions of cure from the constant parameters. The survival of the uncured can be predicted in the flexible parametric cure model in the same way as the survival of the uncured is predicted in the nonmixture cure model. The median survival time of the uncured is predicted using a Newton–Raphson algorithm in a way similar to that of Lambert et al. (2010).

The mixture and nonmixture cure models are sometimes used in situations when cure is not reached within the available follow-up time of the data. This can be done because the models fit an asymptote for the relative survival function. However, estimates of cure can be very sensitive to the parametric distribution chosen. We do not recommend extrapolation in this way when using the flexible parametric cure model, because the point of cure is chosen through the location of the last knot. Even though the position of the last knot can be outside the data, the relative survival should be assumed to flatten within the available follow-up time.

In contrast to the mixture and nonmixture cure model, it is possible to informally test the assumption of a cure proportion for the flexible parametric cure model. The reason is that it is a restricted standard flexible parametric survival model. But these tests should be interpreted with some caution because the comparison is based on the fit over the whole time-scale and not just toward the end where the cure proportion is estimated. By comparing the fit of different flexible parametric cure models where the last knot is placed at different time points, one might evaluate where cure is a reasonable assumption and, from that, get an estimate of the time of cure. Even though this is a quantity of high clinical interest, we do not encourage its estimation from the flexible parametric cure model, because comparison between different models mostly relies on differences at the beginning of the time-scale, where most of the data are.

3 stpm2 command for cure modeling

The `stpm2` command has been updated to enable flexible parametric cure modeling. Only the new options and postestimation commands are discussed here.

3.1 Options

`cure` is used when fitting cure models. It forces the cumulative 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 95th centile. Alternative knot locations can be selected using the `knots()` option. Cure models can only be used when modeling on the log cumulative-hazard scale (`scale(hazard)`).

3.2 Postestimation

`cure` predicts the cure proportion after fitting a cure model.

`uncured` can be used after fitting a cure model. It can be used with the `survival`, `hazard`, and `centile()` options to base predictions for the uncured group.

`startunc(#)` sets the starting value for the Newton–Raphson algorithm for estimating a centile of the survival time distribution of the uncured; the default is the 12.5th centile of the observed follow-up times.

4 Example

To illustrate the `stpm2` command, we use data for 33,874 females aged 50 and over and who were diagnosed with ovarian cancer in England and Wales. The data were obtained from the public-use dataset of all England and Wales cancer registrations between 1 January 1981 and 31 December 1990, with follow-up until 31 December 1995 (Coleman et al. 1999a,b). Background mortality rates were obtained from England and Wales national mortality statistics by age, geographical region, period of diagnosis, and deprivation group (Coleman et al. 1999b). The covariates investigated here are deprivation and age at diagnosis. Deprivation is defined in terms of the area-based Carstairs score and divided into five deprivation categories ranging from the least deprived (affluent) to the most deprived quantile in the population. Age is split into four groups: 50–59, 60–69, 70–79, and 80 and over. This dataset has previously been used to illustrate the `strsmix` and `strsnmix` commands for fitting mixture and nonmixture cure models (Lambert 2007).

4.1 Estimation in one sample

Below are the commands and output from fitting the flexible parametric cure model with `stpm2` to the 50–59 age group.

```
. use ovary_cancer
(Ch31 Adult Ovary 183)
. stset survtime, fail(dead) id(ident)
                id:  ident
     failure event:  dead != 0 & dead < .
obs. time interval:  (survtime[_n-1], survtime]
 exit on or before:  failure
------------------------------------------------------------------------------
    33874  total obs.
        0  exclusions
------------------------------------------------------------------------------
    33874  obs. remaining, representing
    33874  subjects
    28685  failures in single failure-per-subject data
 88539.89  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =    14.992
```

```
. stpm2 if cage==1, scale(hazard) df(5) bhazard(rate) cure
Iteration 0:   log likelihood = -15014.763
Iteration 1:   log likelihood = -14993.781
Iteration 2:   log likelihood = -14993.715
Iteration 3:   log likelihood = -14993.715
Log likelihood = -14993.715                     Number of obs   =       8905
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
_rcs1	1.103296	.0153149	72.04	0.000	1.073279	1.133312
_rcs2	-.20017	.0087823	-22.79	0.000	-.217383	-.1829571
_rcs3	-.0465856	.0059093	-7.88	0.000	-.0581676	-.0350035
_rcs4	-.0019929	.0044742	-0.45	0.656	-.0107622	.0067764
_rcs5	0	(omitted)				
_cons	-.7336044	.0145752	-50.33	0.000	-.7621714	-.7050375

The data are `stset` with the variable `survtime` for the survival time (in years) and the variable `dead` as the event indicator. The `bhazard()` option in the `stpm2` command invokes relative survival models, and the `rate` variable holds the expected mortality rate (hazard) at the time of death. This was obtained from Coleman et al. (1999b) and has previously been merged into the dataset. The `scale(hazard)` option fits a model on the log cumulative-hazard scale, which is the only scale that works for modeling cure within `stpm2`. The `df(5)` option specifies that 5 degrees of freedom are to be used for the restricted cubic spline function, which will place knots at the minimum, maximum, 25th, 50th, 75th, and 95th centiles of the distribution of uncensored survival times.

From the output, we can see that the last spline parameter is omitted. This is the restriction that imposes a constant log cumulative excess hazard after the last knot and therefore a cure point. The constant parameter is the log cumulative excess hazard at and beyond the last knot and can therefore be used to predict cure as $\exp\{-\exp(0.247)\} = 0.278$. Using the `predict` command after fitting a flexible parametric cure model can also provide predictions of the cure proportion as well as the excess-hazard function and the relative survival function for both the sample as a whole and the uncured group. Predictions of the survival time for a given centile of the survival function for the uncured group can also be estimated. The predictions are conditional on any covariates and evaluated at each observed survival time (`_t`). It is also possible to do out-of-sample predictions by using the `timevar()` and `at()` options to specify a covariate pattern and time variable for which to make predictions.

```
. predict rs_all, survival
. predict rs_uncured, survival uncured
. predict cure1, cure
```

Confidence intervals can be obtained for the various predictions by adding the `ci` option to the `predict` command. The standard errors are calculated by using `predictnl`, which uses the delta method. The standard errors for relative survival and cure proportions are obtained on the log(-log) scale (that is, the log cumulative-excess-hazard scale); standard errors for the excess hazard are obtained on the log excess-hazard scale. Figure 1 shows the predicted relative survival for the group as a whole, the predicted relative survival for the uncured, and the predicted cure proportion as a horizontal line.

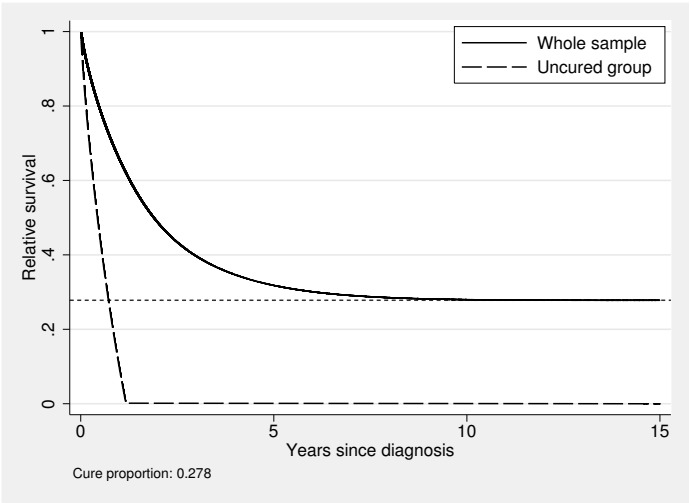

Figure 1. Estimated relative survival

4.2 Modeling the cure proportion

It is often of interest to model the cure proportion; here we include age and deprivation group in the model. Below are the commands and output from fitting a proportional excess-hazards model with cure.

```
. stpm2 cage2-cage4 dep2-dep5, scale(hazard) df(5) bhazard(rate) cure
  (output omitted)
Iteration 3:   log likelihood = -43496.839
Log likelihood = -43496.839                     Number of obs   =      33874
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
cage2	.2638381	.0169397	15.58	0.000	.2306368	.2970393
cage3	.5890439	.0176086	33.45	0.000	.5545316	.6235562
cage4	.9285795	.0219185	42.37	0.000	.8856199	.971539
dep2	.0318505	.0194612	1.64	0.102	-.0062928	.0699938
dep3	.0715027	.0194931	3.67	0.000	.0332968	.1097085
dep4	.0870953	.0198751	4.38	0.000	.0481409	.1260498
dep5	.1294347	.0215438	6.01	0.000	.0872095	.1716598
_rcs1	1.185378	.007485	158.37	0.000	1.170708	1.200048
_rcs2	-.0990392	.004468	-22.17	0.000	-.1077963	-.0902822
_rcs3	.0291592	.0031332	9.31	0.000	.0230182	.0353001
_rcs4	.0390248	.0024252	16.09	0.000	.0342714	.0437781
_rcs5	0	(omitted)				
_cons	-1.111016	.0182878	-60.75	0.000	-1.14686	-1.075173

The estimates for age and deprivation groups are log excess-hazard ratios, so the excess hazard for the oldest age group is $\exp(0.929) = 2.53$ times higher than the excess hazard for the youngest group. The flexible parametric cure model is a special case of a nonmixture cure model; the excess-hazard ratios from this model are very similar to those from a nonmixture cure model with a log(-log) link (Lambert 2007). The parameter estimates are also transformations of cure. For example, the cure proportion for the youngest age group in the least deprived group is $\exp\{-\exp(0.087)\} = 0.34$ and for the oldest age group in the least deprived group is $\exp\{-\exp(0.087+0.929)\} = 0.063$.

Nonproportional hazards are common in population-based cancer studies and can be modeled by including interactions between covariates and splines for time. To let the effects of covariates in the flexible parametric cure model vary by time is equivalent to modeling the parameters in the parametric distribution used in a nonmixture cure model. It has been shown for nonmixture cure models where a Weibull distribution is used that modeling of both Weibull parameters can be crucial (Lambert et al. 2007b). Similarly, we believe that time-dependent effects should usually be included in the flexible parametric cure model for most cancers. Below are the commands and output from fitting the flexible parametric cure model with time-dependent effects for both age and deprivation group.

```
. stpm2 cage2-cage4 dep2-dep5, scale(hazard) df(5) knscale(time) bhazard(rate)
> cure tvc(cage2-cage4 dep2-dep5) dftvc(3)
```

(*output omitted*)

```
Iteration 4:   log likelihood = -42861.054
Log likelihood = -42861.054                     Number of obs   =      33874
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
xb						
cage2	.4198831	.0233391	17.99	0.000	.3741394	.4656268
cage3	.8343296	.0230064	36.27	0.000	.7892379	.8794213
cage4	1.174375	.0263022	44.65	0.000	1.122824	1.225926
dep2	.0414502	.0227631	1.82	0.069	-.0031646	.086065
dep3	.0999266	.0225892	4.42	0.000	.0556526	.1442007
dep4	.1353842	.0228843	5.92	0.000	.0905318	.1802365
dep5	.195961	.0244765	8.01	0.000	.147988	.2439341
_rcs1	1.446098	.0276676	52.27	0.000	1.39187	1.500325
_rcs2	-.235717	.0142901	-16.50	0.000	-.2637252	-.2077089
_rcs3	-.0690412	.0073471	-9.40	0.000	-.0834411	-.0546412
_rcs4	.0020715	.0031142	0.67	0.506	-.0040323	.0081752
_rcs5	0	(omitted)				
_rcs_cage21	-.1472517	.0275937	-5.34	0.000	-.2013344	-.093169
_rcs_cage22	.098554	.0144949	6.80	0.000	.0701445	.1269635
_rcs_cage23	0	(omitted)				
_rcs_cage31	-.2566918	.0265035	-9.69	0.000	-.3086376	-.2047459
_rcs_cage32	.2333217	.0141668	16.47	0.000	.2055552	.2610881
_rcs_cage33	0	(omitted)				
_rcs_cage41	-.3920616	.027467	-14.27	0.000	-.445896	-.3382273
_rcs_cage42	.329778	.0154273	21.38	0.000	.2995411	.3600148
_rcs_cage43	0	(omitted)				
_rcs_dep21	-.0124399	.0244274	-0.51	0.611	-.0603167	.0354369
_rcs_dep22	.0043985	.014359	0.31	0.759	-.0237447	.0325416
_rcs_dep23	0	(omitted)				
_rcs_dep31	-.0430759	.0238987	-1.80	0.071	-.0899164	.0037647
_rcs_dep32	.0186057	.0140738	1.32	0.186	-.0089784	.0461898
_rcs_dep33	0	(omitted)				
_rcs_dep41	-.0631187	.0240329	-2.63	0.009	-.1102222	-.0160151
_rcs_dep42	.0471661	.0141666	3.33	0.001	.0194001	.0749321
_rcs_dep43	0	(omitted)				
_rcs_dep51	-.0962763	.0251416	-3.83	0.000	-.1455529	-.0469997
_rcs_dep52	.0697525	.0150006	4.65	0.000	.0403518	.0991532
_rcs_dep53	0	(omitted)				
_cons	-1.353309	.0238595	-56.72	0.000	-1.400073	-1.306546

The estimates for age and deprivation groups in the table above are harder to interpret: they are no longer log excess-hazard ratios, but they are still transformations of cure. Below are the predicted cure proportions and median survival times of the uncured for each calendar period and deprivation group.

```
. predict cure, cure ci
. predict med_survunc, centile(50) uncured ci
. by cage caquint, sort: generate first= _n == 1
. tabdisp cage caquint if first, cellvar(cure med_survunc) format(%5.3fc)
```

Age Group	GB quintile Carstairs score leastdep	2	3	4	mostdep
50-59	0.772	0.764	0.752	0.744	0.730
60-69	0.675	0.664	0.648	0.637	0.620
70-79	0.551	0.538	0.518	0.506	0.485
80+	0.433	0.418	0.397	0.384	0.362

The results look similar to results from a corresponding nonmixture cure model (see Lambert [2007]), but there is a tendency of lower estimated cure proportions for the oldest age group, where we know that the nonmixture cure model can overestimate cure.

4.3 Sensitivity to knot placement

The flexible parametric survival model has been shown to be robust to the number and location of the knots (Nelson et al. 2007; Lambert and Royston 2009). To evaluate the sensitivity to the location of the knots for the flexible parametric cure model, we compared the predicted survival and cure proportion from the flexible parametric cure model with different knot positions and numbers of knots using the age group 70–79. The fit of the cure model is fairly robust to the number and location of the knots (figure 2).

For a standard `stpm2` model, the default positions for the knots are distributed evenly according to centiles of the uncensored event times. When the default knot positions from a standard `stpm2` are used, all knots except the last one are placed within the first few years from diagnosis because most of the events happen early on and cure (where the survival reaches a plateau) seems slightly overestimated. Cure is also slightly overestimated when the last knot is placed at the 95th centile of death times. The number of knots seems to have little impact on the estimated relative survival.

Overall, the flexible parametric cure model seems to give a good fit, so long as knots are placed over the whole follow-up period and the last knot is positioned at the last observed death time or possibly later. When the `cure` option is used with the `stpm2` command, the default knot positions are according to centiles of the uncensored event times except for one knot, which is placed at the 95th centile.

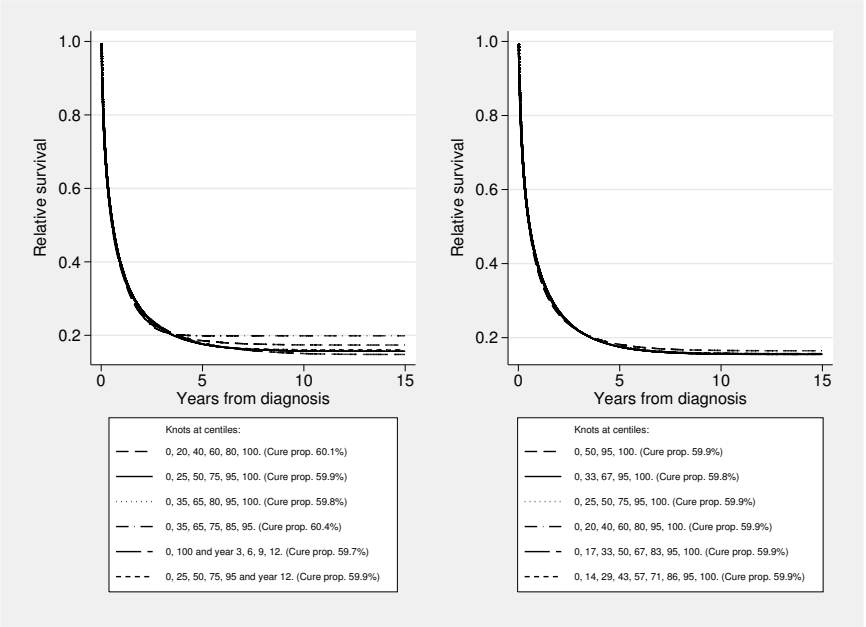

Figure 2. Predicted relative survival from flexible parametric cure models with different knot locations and numbers of knots

4.4 Comparison of flexible parametric cure model and nonmixture cure model

It has been shown that the mixture and nonmixture cure models can give biased estimates of cure for older age groups because of a very high excess hazard within the first months after diagnosis that is not captured by the parametric distributions often used. Lambert (2007) has suggested alternative approaches, one using a mixture of Weibull distributions and one splitting the time-scale in two parts and estimating the excess hazard separately in the two intervals. These approaches have some limitations: the split-time model needs a subjectively chosen cutpoint and leads to a noncontinuous excess-hazard function, and there are sometimes problems with convergence for the mixture of Weibull distributions. The flexible parametric cure model can in a less complicated way overcome the problem with the standard parametric cure models because the splines can more easily capture the underlying shape of the survival distribution. This is illustrated here using the ovarian cancer data restricted to the oldest age group (80 years and older at diagnosis), with no modeling of covariates.

Figure 3 shows the predicted relative survival from the Weibull nonmixture cure model and the flexible parametric cure model along with empirical life-table estimates of relative survival. As with other cure models, the flexible parametric cure model will give an estimate of the cure proportion even when cure is not reasonable. It is therefore important to always compare results from cure models with empirical estimates of relative survival and to make sure that there seems to be a proportion of patients who are cured. This is not a specific drawback for the flexible parametric cure model but for cure models in general. Figure 3 shows that the Weibull nonmixture cure model seems to overestimate cure and underestimate survival for the first two years of follow-up. The flexible parametric cure model follows the life-table estimates more closely and gives a lower, less biased estimate of cure. There is still discrepancy between the flexible parametric cure model and the life-table estimates toward the end of follow-up, but the data are sparse, and an even better fit could be achieved by adding an extra knot.

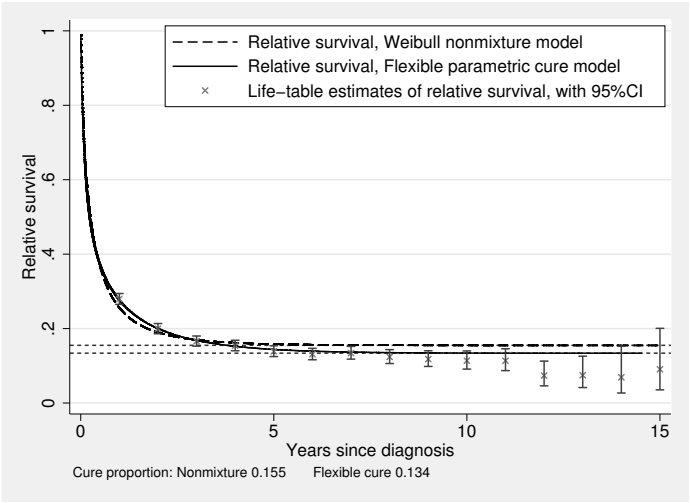

Figure 3. Predicted relative survival from the flexible parametric cure model and Weibull nonmixture cure model, along with life-table estimates of relative survival

5 Conclusion

Cure models within the framework of flexible parametric survival models enable cure modeling when standard models are not flexible enough. The `stpm2` command for fitting flexible parametric survival models has been updated to incorporate this extension.

6 Acknowledgments

We thank Paul Dickman and Sandra Eloranta for their help with developing the flexible parametric cure model and for valuable comments on this article.

7 References

Andersson, T. M.-L., 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: 96.

Begg, C. B., and D. Schrag. 2002. Attribution of deaths following cancer treatment. *Journal of the National Cancer Institute* 94: 1044–1045.

Coleman, M. P., P. Babb, P. Damiecki, P. Grosclaude, S. Honjo, J. Jones, G. Knerer, A. Pitard, M. J. Quinn, A. Sloggett, and B. De Stavola. 1999a. Cancer survival trends in England and Wales, 1971–1995: Deprivation and NHS Region. Office for National Statistics, London, UK.

Coleman, M. P., P. Babb, D. Mayer, M. J. Quinn, and A. Sloggett. 1999b. Cancer survival trends in England and Wales, 1971–1995: Deprivation and NHS Region. CD-ROM. London, UK: Office for National Statistics.

De Angelis, R., R. Capocaccia, T. Hakulinen, B. Soderman, and A. Verdecchia. 1997. Mixture models for cancer survival analysis: Application to population-based data with covariates. *Statistics in Medicine* 18: 441–454.

Dickman, P. W., and H. O. Adami. 2006. Interpreting trends in cancer patient survival. *Journal of Internal Medicine* 260: 103–117.

Lambert, P. C. 2007. Modeling of the cure fraction in survival studies. *Stata Journal* 7: 351–375.

Lambert, P. C., P. W. Dickman, P. Österlund, T. Andersson, R. Sankila, and B. Glimelius. 2007a. Temporal trends in the proportion cured for cancer of the colon and rectum: A population-based study using data from the Finnish Cancer Registry. *International Journal of Cancer* 121: 2052–2059.

Lambert, P. C., P. W. Dickman, C. L. Weston, and J. R. Thompson. 2010. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *Journal of the Royal Statistical Society, Series C* 59: 35–55.

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

Lambert, P. C., J. R. Thompson, C. L. Weston, and P. W. Dickman. 2007b. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 8: 576–594.

Nelson, C. P., P. C. Lambert, I. B. Squire, and D. R. Jones. 2007. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 26: 5486–5498.

Royston, P. 2001. Flexible parametric alternatives to the Cox model, and more. *Stata Journal* 1: 1–28.

Royston, P., and M. K. B. 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: 2175–2197.

Verdecchia, A., R. De Angelis, R. Capocaccia, M. Sant, A. Micheli, G. Gatta, and F. Berrino. 1998. The cure for colon cancer: Results from the EUROCARE study. *International Journal of Cancer* 77: 322–329.

Yu, B., R. C. Tiwari, K. A. Cronin, and E. J. Feuer. 2004. Cure fraction estimation from the mixture cure models for grouped survival data. *Statistics in Medicine* 23: 1733–1747.

About the authors

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

Paul C. Lambert is a reader in medical statistics 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.