Cancer survival: principles, methods and applications
Exercises on secondary measures

Paul W. Dickman

February 2023

Contents

1 Downloading user-written Stata commands and data files

1.1 Downloading the course files

The course files (e.g., data files and solution do files) are distributed as a Stata package so should be downloaded from within Stata. It is suggested that you create a new directory, change the Stata working directory to the new directory (e.g., `cd c:\survival\`), and then download the files. You can create a new directory in Windows Explorer or you can do it from within Stata as follows.

```
mkdir c:\survival
cd c:\survival
```

Use the `pwd` command to confirm you are in the working directory you wish to use for the course and then issue the following command from the Stata command line to install the course files.

```
net install http://www.pauldickman.com/survival/secondary_measures, all replace
```

`net install` downloads the files and copies them to appropriate directories according to the way Stata is setup. Ancillary files (e.g., PDF, XLS, DTA) are copied to the current working directory; ADO and HLP files are installed into the appropriate directory according to the way Stata is configured.

1.2 Installing Stata user-written commands for relative survival

Standard Stata does not contain any commands for estimating and modelling relative survival so we must extend Stata using commands written by users. Download and installation is done within Stata. It is recommended that you change the Stata working directory to the course directory (e.g., `cd c:\survival\`) before issuing these commands.

1.2.1 How can I check if these commands are already installed?

You can use the `which` command to check if (and where) a Stata command is installed.

```
. which stpm2
c:\ado\plus\s\stpm2.ado
*! version 1.7.6 18Jan2023
```

Use the `adoupdate` command to update previously installed user-written commands (note that this is distinct from the `update` command that updates official Stata commands). Simply type `adoupdate, update` to update all user-written commands.

1.2.2 strs - estimating and modelling relative survival

The `strs` command, written by Paul Dickman and Enzo Coviello can be downloaded by typing the following:

```
. net install http://www.pauldickman.com/rsmodel/stata_colon/strs, all replace
```

Note that some of the data files are contained in both the `strs` and the `course_files` packages, hence the need for the `replace` option. See `https://pauldickman.com/software/strs/` for further details about the command or read the Stata help file after installation. The command is described in a Stata Journal article [1].

1.2.3 stpm2 - flexible parametric models

The `stpm2` command, written by Paul Lambert and Patrick Royston, fits flexible parametric survival models (so called Royston-Parmar models). Relative survival models can be fitted using the `bhazard()` option. It is installed from within Stata using the following commands:

```
ssc install stpm2
ssc install rcsgen
```

The command is described in a Stata Journal article [2]. `rcsgen` is a command for generating basis vectors for restricted cubic splines and is required by `stpm2`. Flexible parametric cure models (fitted using an option to `stpm2`) are described in another Stata Journal article [3].

Further details at `https://pclambert.net/software/stpm2/`

1.2.4 standsurv - standardized survival and related functions

The `standsurv` command, written by Paul Lambert, estimates standardized survival curves and related measures. It also allows various contrasts between the standardized functions. It is a post-estimation command and can be used after fitting a wide range of survival models, including `streg` (except generalized gamma), `stpm2` and `strcs`. It is installed from within Stata:

```
ssc install standsurv
```

Further details at `https://pclambert.net/software/standsurv/`

1.2.5 strsmix and strsnmix - cure models

To install `strsmix` and `strsnmix` (commands for fitting cure models) first type `findit lambert cure` then click on the Stata Journal link followed by *click to install*. These commands are described in a Stata Journal article [4].

1.2.6 Estimating probability of death in a competing risks framework

The `stcompet` command estimates the cumulative incidence function (CIF) non-parametrically. The `stcompadj` command estimates the CIF using a competing risks analogue of the Cox model. The `stpm2cm` command estimates the crude probabilities of death (i.e., CIF) after fitting a relative survival model using `stpm2`. The `stpm2cif` command estimates the CIF through postestimation after fitting a cause-specific competing risks model using `stpm2`.

```
ssc install stcompet
ssc install stcompadj
ssc install stpm2cm
ssc install stpm2cif
```

The `stpm2cif` command is described in a Stata Journal article [5].

2 Exercises

244. **Age standardization using flexible parametric models (external standard)**

This question uses stpm2 to obtain model-based age-standardized estimates of relative survival. The analytic approach is similar to exercise ?? where we used a model-based approach with an internal standard (that is, we standardised to the age distribution of the patient cohort). We will now age-standardise using an external standard population, namely the International Cancer Survival Standard (ICSS) [6]. The estimates should be similar to those obtained in exercise ?? where we also age-standardised using ICSS weights, but used a nonparametric (life-table) rather than a model-based approach.

(a) Run the code in q244.do up to and including tab agegroup to read the data, merge in the external rates, and create variables for analysis. Ensure you understand what each code segment does (ask if you are uncertain) and then look at the age distribution.

```
. tab agegrp, gen(agegrp)

  Age group |      Freq.     Percent        Cum.
------------+-----------------------------------
      15-44 |        647       30.23       30.23
      45-54 |        397       18.55       48.79
      55-64 |        464       21.68       70.47
      65-74 |        401       18.74       89.21
        75+ |        231       10.79      100.00
------------+-----------------------------------
      Total |      2,140      100.00
```

Exercise 243 contains a table of International Cancer Survival Standard (ICSS) weights for broad age groups. Identify the appropriate weights for melanoma and compare the ICSS weights to age distribution of the melanoma patients (i.e., the table above). Based on this comparison and the knowledge that net survival declines with increasing age at diagnosis, how to you expect the age-standardised net survival to differ from the crude (not age standardised) net survival?

(b) We will now create an individual weight for each observation, calculated as the proportion of patients in the given age group in the standard population (ICSSwt) divided by the proportion in that age group in our patient data (a_age). Age groups which are under-represented in our data (compared to the standard population) will receive a weight greater than one (and vice versa).

```
recode agegrp (1=0.28) (2=0.17) (3=0.21) (4=0.20) (5=0.14), gen(ICSSwt)
local total= _N
bysort agegrp: generate a_age = _N/‘total’
generate w = ICSSwt/a_age
```

_N is a Stata system variable containing the total number of observations in the dataset. However, when we reference _N within bysort its value will be the number of observations within the by group. In the code above we write the total number of observations to a local macro variable (total). In this way, we generate the variable a_age to contain the proportion of observations in each age group.

(c) Let's have a look at the values of the weights.

```
. tabstat w, statistics(mean min max) by(agegrp)

agegrp |      mean       min       max
-------+------------------------------
 15-44 |  .9261205  .9261205  .9261205
 45-54 |  .9163728  .9163728  .9163728
 55-64 |  .9685345  .9685345  .9685345
 65-74 |  1.067332  1.067332  1.067332
   75+ |  1.29697   1.29697   1.29697
--------------------------------------
```

We see that the weights are constant within each age group and that patients aged 45–54 will be slightly down-weighted (18.5% of our cohort compared to 17% in the standard population) whereas patients in age group 65–74 will be slightly up-weighted (18.7% of our cohort compared to 20% in the standard population). The age distribution of our cohort is slightly younger than that of the standard population, so we expect the age-standardised estimates to be slightly lower than the crude estimates.

(d) We have created the weights, but will not use them just yet. We now fit a flexible parametric model adjusting for calendar year of diagnosis (as a restricted cubic spline) and age group.

```
. stpm2 agegrp2 agegrp3 agegrp4 agegrp5, scale(h) df(4) eform ///
>   tvc(agegrp2 agegrp3 agegrp4 agegrp5) dftvc(1) bhazard(rate)
```

We will now predict the marginal (population-averaged) survival function for the cohort, both with and without standardisation.

We first create a temporary time variable (`temptime`) so as to predict survival for each of 101 unique values of time (every 0.1 years from 0 to 10) rather than for each of the 5,308 observations in the data set.

```
. range temptime 0 10 101

. // marginal (population-averaged) unstandardised survival
. predict s_unweighted, meansurv timevar(temptime)

. // marginal (population-averaged) survival standardised to ICSS
. predict s_weighted, meansurv timevar(temptime) meansurvwt(w)
```

We start with predicting the unweighted (unstandardised) marginal survival. At each of the values of time, we predict the marginal (population-averaged) survival using the `meansurv` option to the predict command. For time=0, the marginal survival is trivially 1. The next value of `temptime` is 0.1. What `meansurv` effectively does is predict the survival at time 0.1 for each of the 5,308 observations in the data set (conditional on covariates for each observation) and then takes the average of these 5,308 observations. This continues for each of the other values of `temptime`.

Another way of conceptualising this, is that for each of the 5,308 observations in the data set we predict the survivor function (from time 0 to 10) for the individual with given values of year and age. We then average the 5,308 individual survival curves to get the marginal survival curve. This is stored in the variable `s_unweighted`.

To get the age-standardised marginal (population-averaged) net survival, we use the same procedure but apply weights when averaging the individual survival curves. The age-standardised estimates are stored in `s_weighted`. Predicting and averaging a large number of individual survival curves can be computationally intensive.

(e) Compare the values of `s_unweighted` and `s_unweighted` at 1, 5, and 10 years. Before looking at the estimates, estimate the magnitude of the difference you expect to see.

```
. list temptime s_unweighted s_weighted if inlist(temptime, 1, 5, 10)
```

There is code in `q244.do` to graph the standardised and unstandardised survival curves.

(f) Repeat the exercise using the colon cancer data; you will need to change the ICSS weights. The colon cancer patients are older than the standard population so you will see differences between the age-standardised and unstandardised estimates.

250. **Probability of death in a competing risks framework (life table relative survival)**

`strs` implements the approach proposed by Cronin and Feuer (2000) [7] for estimating the crude probability of death based on life table estimates of relative survival. We explore the life table approach in this question. Lambert *et al.* (2010) [8] subsequently showed how the estimates can be obtained after fitting a relative survival model, namely a flexible parametric models for relative survival, which use restricted cubic splines for the baseline cumulative excess hazard and for any time-dependent effects. The approach using flexible parametric models for relative survival is covered in question 251. Although the two approaches estimate the same quantity, the life table approach provides estimates for grouped data so we get an estimated probability for an age group rather than an estimate for a specific age as can be obtained in the model-based approach.

(a) Load the Melanoma data, drop subjects diagnosed 1975–1984 and then and use `strs` to obtain life-tables stratified by age group and sex. Use the `cuminc` option to obtain the crude probabilities of death due to cancer and due to other causes.

(b) How is the probability of death due to all causes, `F`, calculated?

(c) Why is the crude probability of death due to cancer, `ci_dc` similar to the all-cause probability of death for subjects aged 0-44?

(d) For both males and females aged 60-74 what is the probability of death due to all-causes at 5 years post diagnosis? What two variables can be added together to give the probability of death due to all-causes?

(e) What proportion of the all-cause deaths at 5 years post diagnosis are due to cancer and due to other causes for males? Compare these figures for the different age groups.

(f) The age groups are fairly wide, explain how you would expect the crude probability of death due to cancer to differ between a 60 and 74 year old, even if the relative survival was identical.

(g) Plot the net probability of death, the crude probability of death due to cancer and the overall probability of death for males by age group. Try to understand the relationship between these various measures.

251. **Probability of death in a competing risks framework (relative survival model)**

In exercise 250 we explored how one could estimate crude probabilities of death based on life table estimates of relative survival making use of the `strs` implementation of the approach proposed by Cronin and Feuer (2000) [7]. Lambert *et al.* (2010) [8] subsequently showed how the estimates can be obtained after fitting a relative survival model, namely a flexible parametric models for relative survival, which use restricted cubic splines for the baseline cumulative excess hazard and for any time-dependent effects. Although the two approaches estimate the same quantity, the life table approach provides estimates for grouped data so we get an estimated probability for an age group rather than an estimate for a specific age as can be obtained in the model-based approach.

(a) Load the Melanoma data and merge in the background mortality rates as in question **??**. Fit a flexible parametric relative survival model including age group with time-dependent effects.

```
. tab agegrp, gen(agegrp)
. stpm2 agegrp2-agegrp4, scale(hazard) bhazard(rate) df(5) ///
      tvc(agegrp2-agegrp4) dftvc(2)
```

Calculate the estimated net probability of death (1 - relative survival) and plot the four curves on a single graph. Interpret the plot.

(b) Use the `standsurv` command to estimate the crude probability of death. Note that `standsurv` will predict for individual covariate patterns and for specific ages at diagnosis. Perform the predictions for males aged 40, 55, 70 and 80 diagnosed in 1985. As we are only making predictions for one individual, we need to create a variable with age at diagnosis and date at diagnosis for the healthy individual to match to. This is used as the prediction for the expected survival. We also define a user-defined mata function `calc_allcause` to calculate the all-cause failure function as a sum of the two `at()` options. The prediction for a 40 year old (the first age group) can be obtained using,

```
. mata function calc_allcause(at) return(at[1]+at[2])

. range temptime2 0 5 101
. gen aged = .
. gen dated = mdy(1,1,1985) in 1
. replace aged = 40 in 1

. standsurv if _n==1,  at1(sex 1 agegrp2 0 agegrp3 0 agegrp4 0) ///
>  verbose timevar(temptime2) ///
>  atvar(crprob1) crudeprob stub2(cancer other) ///
>  expsurv(using("Z:\cansurv\data\popmort.dta") ///
>          datediag(dated)               ///
>          agediag(aged)               ///
>          pmrate(rate)                       ///
>          pmage(_age)                                ///
>          pmyear(_year)              ///
>          pmother(sex)                       ///
>          pmmaxyear(1985)                 ///
>          at1(sex 1))                                        ///
>  userfunction(calc_allcause) ///
>  userfunctionvar(allcause1) transform(none)
```

Plot the estimated crude probability of death due to cancer for each of the selected ages on the same graph. Contrast these with the estimated net probability of death from part (a).

(c) Generate a similar plot but for the crude probability of death due to other causes.

(d) A useful way of presenting crude probabilities is through stacked graphs.

   i. Generate the stacked graphs for each of the selected ages. Use the solution Do file for help.

   ii. Now overlay the net probability of death. Does it better illustrate the contrast described in (b)?

(e) Advanced: Now fit a model using splines for the effect age with the spline terms allowed to be time-dependent.

   i. Calculate the crude probabilities of death and compare these to the model where age is categorized.

   ii. Now calculate crude probabilities of death at individual ages from 40 to 90 years old at 5 years since diagnosis - plot these over age. See do file for help. Hint: you will need to do a loop over 50 `standsurv` predictions.

260. **Fitting cure models**

Stata addon required! This exercise requires the Stata user-written command `strsmix`. See Section 1.2 (page 2) for details and installation instructions.

We will now apply cure fraction models [9, 10] to the colon cancer data. In this exercise we fit mixture cure models and in exercise 261 we fit flexible parametric cure models. The cure fraction models treat time as continuous and thus there is no need to split the time scale. However, the expected hazard (mortality) rate at the time of death (or censoring) is required. Use the following commands to merge in the expected mortality rate.

```
. use colon
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120)
. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort,  keep(match master)
```

The `scale(12)` option converts survival time to years. The `exit(time 120.5)` option creates a maximum follow-up time of 10 years (120 months).

(a) Explain the purpose of the two `gen` statements in the above stata code.

(b) Fit a mixture cure fraction model to those diagnosed between 1975-1984 using the following command.

```
. strsmix if year8594==0, dist(weibull) link(identity) bhazard(rate)
```

i. What is the estimate of the cure fraction?

Use the following commands to obtain prediction of the relative survival curve and the survival distribution of the 'uncured' and then plot these estimates against time (`_t`)

```
. predict rs7584, survival
. predict rs7584u, survival uncured
```

ii. Does the relative survival curve appear to reach a plateau at the cure fraction? Would you expect it to?

iii. Approximately what proportion of the 'uncured' group have died after 2 years?

iv. Approximately what is the median survival time of the 'uncured'?

(c) Repeat the above for those diagnosed between 1985–1994. Contrast the estimates for the two time periods.

(d) Now we will compare the two time periods more formally by including (`year8594`) as a covariate. First just allow the cure fraction to vary between time periods.

```
. strsmix year8594, dist(weibull) link(identity) bhazard(rate)
```

i. What is the estimated difference in the cure fraction between the two time periods? Contrast this to the estimates obtained in b(i) and (c).

ii. This model is making a fairly strong assumption regarding the survival distribution of the 'uncured' for the two periods. What is this assumption?

Now allow the two Weibull parameters (λ and γ) to vary between the two time periods.

```
. strsmix year8594, dist(weibull) link(identity) bhazard(rate)
        k1(year8594) k2(year8594)
```

iii. What is the estimated difference in the cure fraction between the two time periods? Contrast this with d(i).

iv. Test the assumption that the survival distribution of the 'uncured' is the same for the two time periods.

(e) Now fit a model including age group and time period of diagnosis using a logit link (use option `link(logit)`).

i. Interpret the parameter estimates (you may want to display the exponentiated coefficents bys using `strsmix, eform`).

ii. Obtain predictions of the median survival of the 'uncured'.
Hint, use `predict med, centile` to obtain predicted values of the median.

261. **Fitting cure models using flexible parametric survival models**

Stata addon required! This exercise requires the Stata user-written command `stpm2`. See Section 1.2 (page 2) for details and installation instructions.

We will now apply flexible parametric cure models to the same data as in exercise 260, where we fitted mixture cure models. Read in the data, stset and merge on expected mortality rates in the same way as in exercise 260.

(a) Compare the cure proportion in the two time periods by including the variable `year8594` as a covariate in the stpm2 command. Assume proportional hazards.

```
. stpm2 year8594, df(6) bhazard(rate) scale(hazard) cure
```

i. How do you interpret the coefficient for the effect of the time period?

ii. Use the coefficients in the output to calculate the estimated cure proportions for the two time periods.

iii. Predict the cure proportions using the predict command to check your calculations.

```
. predict cure1, cure
. list cure1 if year8594==0, constant
. list cure1 if year8594==1, constant
```

iv. What is the estimated difference in the cure proportion between the two time periods? Compare this to the estimates obtained in exercise 260. Are the results similar? Would you expect them to be similar?

v. Predict the median survival time of uncured. Is the median survival time the same in the two groups? Should it be?

```
. predict med1, centile(50) uncured
. list med1 if year8594==0, constant
. list med1 if year8594==1, constant
```

(b) Now allow time-dependent effect.

```
. stpm2 year8594, df(6) tvc(year8594) dftvc(4) bhazard(rate) scale(hazard) cure
```

i. How do you interpret the coefficient for the effect of the time period?

ii. Use the coefficients in the output to calculate the estimated cure proportions for the two time periods.

iii. Predict the cure proportions using the predict command to check your calculations.

```
. predict cure2, cure
. list cure2 if year8594==0, constant
. list cure2 if year8594==1, constant
```

iv. Are the cure proportions similar to what was estimated in (a)?

v. Predict the median survival time of uncured. Is the median survival time the same in the two groups? Should it be? Is the difference between the periods smaller or larger than in (a)? Why?

```
. predict med2, centile(50) uncured
. list med2 if year8594==0, constant
. list med2 if year8594==1, constant
```

(c) Plot the estimated overall relative survival and the relative survival among uncured for the two periods. Do the survival curves reach a plateau? Should they?

270. Conditional survival

In this tutorial, we illustrate how to estimate net survival conditional on surviving some time since diagnosis. We estimate conditional net survival using both a non-parametric (life table) approach and based on a flexible parametric model.

The term 'conditional survival' is sometimes used to mean 'conditional on covariates' (to distinguish from marginal survival) and sometimes used to mean 'conditional on having survived up to some time s'. Here we estimate the latter.

The conditional survival function $\text{CS}(t|s)$ is defined as the probability of surviving an additional t years given a patient has already survived s years.

$$\text{CS}(t|s) = P(T > t + s|T > s) = \frac{S(s+t)}{S(s)}$$

We will estimate conditional net survival (CNS),

$$\text{CNS}(t|s) = \frac{S_N(s+t)}{S_N(s)}$$

where $S_N(t)$ is net survival. We will estimate net survival 5 years post diagnosis, conditional on having survived 1 year, using 3 approaches:

(a) Non-parametric (Pohar Perme estimator) by taking the ratio of $S(5)$ to $S(1)$ in a standard cohort life table.

(b) Non-parametric (Pohar Perme estimator) by restricting the cohort to patients who survive 1 year (i.e., late entry)

(c) By predicting the ratio of $S(5)$ to $S(1)$ based on a flexible parametric model

Care must be taken when reporting estimates of conditional survival. We will estimate 'net survival 5 years post diagnosis conditional on having survived 1 year', which is denoted by $\text{CNS}(4|1)$ (the probability of surviving 4 additional years conditional on surviving 1 year). One can see reports presenting 'conditional 5-year survival' and it is not clear if this represents 5 years in addition to s or survival up to 5 years post diagnosis conditional on survival to s.

We have chosen to use the same notation as Belot *et al.* (2019) [11]. References to the use of conditional survival can be found in their tutorial paper.

We will use the colon cancer data for this example, as the principles are best demonstrated using an example with relatively high early mortality.

(a) Tabulate a standard cohort life table with 1-year intervals (using `strs` with the `pohar` and `ht` options option) and divide the 5-year net survival by the 1-year net survival.

```
. use colon.dta, clear

. stset exit, origin(dx) fail(status==1 2) id(id) scale(365.24)

. strs using popmort, br(0(1)10) mergeby(_year sex _age) pohar ht ///
       list(n d y cns_pp lo_cns_pp hi_cns_pp)
```

What is the estimated net survival 5 years post diagnosis conditional on having survived 1 year?

We used the Pohar Perme estimates, but any estimates can be used. The disadvantage of this approach is that we don't get the standard error.

By default, `strs` uses the actuarial approach for estimation (i.e., estimation is performed on the survival scale) but if late entry is detected it estimates the cumulative hazard and then transforms to the survival scale. We will use late entry in the next step, so specify the `ht` option (hazard transformation) to force the same approach to estimation here.

(b) We now restrict the cohort to individuals who survived at least 1 year, by specifying the enter option to `stset`, and tabulate the life table (with the exact same call to `strs`).

```
. stset exit, origin(dx) enter(time dx+365.24) fail(status==1 2) id(id) scale(365.24)

. strs using popmort, br(0(1)10) mergeby(_year sex _age) pohar ht ///
       list(n d y cns_pp lo_cns_pp hi_cns_pp)
```

What is the estimated net survival 5 years post diagnosis conditional on having survived 1 year? What is the 95% confidence interval?

(c) We now use a model-based approach. As with the first life table approach, we estimate survival from diagnosis and divide the estimated 5-year survival by the estimated 1-year survival. As we are modelling net survival, we need to merge in the external rates.

```
. // Return to original stset (everyone at risk from diagnosis)
. stset exit, origin(dx) fail(status==1 2) id(id) scale(365.24)

. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort, keep(match master)
.
. // Fit the model without covariates
. stpm2, scale(hazard) df(5) bhazard(rate)
```

Rather than just estimate CNS(4|1) we'll estimate CNS(t|1) for a range of values of t. We'll first predict the unconditional relative survival at each value of `_t` and store the estimates in a new variable `s`.

Next we want to predict $S(t)/S(1)$ for a range of values of t (where t is now time since diagnosis).

We will create two temporary time variables; `timevar` will take 100 values between 1 and 5 while `t1` will be set to 1 for each of the observations. We then predict the ratio of S(timevar) / S(t1).

```
. predict s, survival ci

. range timevar 1 5 100

. gen t1 = 1 in 1/100

. predictnl condsurv = ///
   predict(survival timevar(timevar)) / predict(survival timevar(t1))
```

What is the estimated net survival 5 years post diagnosis conditional on having survived 1 year?

Note that we didn't request confidence intervals, which we will do now. We will predict on the log scale in order to get more appropriate confidence intervals.

```
. predictnl condsurv1 = ln(predict(survival timevar(timevar)) / ///
>           predict(survival timevar(t1))) , ///
>           ci(condsurv1_lci condsurv1_uci)
```

How does the estimated conditional survival (and confidence intervals) compared to those you obtained in the previous part from life tables?

(d) Now plot the conditional survival as a function of time (see the code in `q270.do`). The do file also contains code for estimating and plotting unconditional and conditional survival on the same graph.

282. **Calculating excess and 'avoidable' deaths from life tables**

(a) Load the Melanoma data, drop subjects diagnosed 1975-1984 and then and use `strs` to obtain life-tables stratified by age group and sex. Load the grouped data and keep the following variables.

```
. keep start end n cp cp_e2 cr_e2 sex agegrp
```

(b) What is the difference in five-year relative survival between males and females in each age group?

(c) We will now investigate excess deaths and 'avoidable' deaths. The question of interest is how many fewer deaths we would expect to see if males could achieve the same relative survival as females. To do this we will reshape the data from long form to wide form to make calculations easier.

```
. bysort sex (agegrp start): gen j = _n
. gen sexlab =cond(sex==1,"_m","_f")
. drop sex
. reshape wide start end n cp cp_e2 cr_e2 agegrp, i(j) j(sexlab) string
. rename agegrp_m agegrp
. rename start_m start
. rename end_m end
. drop agegrp_f start_f end_f
```

Look at the data in the data browser to make sure you understand what the `reshape` command has done.

(d) In order to calculate the predicted number of deaths we need to define how many subjects were at risk at the the start of follow-up. For simplicity, we will use the average number of cases per year over the 10 year diagnosis period. This can be calculated as follows.

```
. bys agegrp: gen Nrisk_m = n_m[1]/10
```

Calculate the overall (all-cause) probability of death, $1 - S^*(t)R(t)$, for males.

```
. gen p_dead_m = 1 - cp_e2_m * cr_e2_m
```

For males, calculate the expected number of all-cause deaths, `Nd_m`, the expected number of deaths if the study population were free of cancer, `NExp_d_m` and the excess deaths associated with a diagnosis of cancer, `ED_m`.

```
. gen Nd_m = Nrisk_m*p_dead_m
. gen NExp_d_m = Nrisk_m*(1-cp_e2_m)
. gen ED_m = Nd_m - NExp_d_m
```

i. How many all cause deaths would we expect to see in each age group at 5 years post diagnosis?

ii. How many more deaths are there than would be expected in a similar cancer free group in the population?

iii. How many excess deaths by 5 years are associated with a diagnosis of melanoma over all age groups?

(e) Repeat the above calculations for females. How do the excess deaths for females compare to the males?

(f) We will now apply the relative survival estimates for females to the males' expected survival in order to calculate the 'avoidable' deaths.

```
. gen Nd_m_f = Nrisk_m*(1 - cp_e2_m * cr_e2_f)
. gen AD_m = Nd_m - Nd_m_f
```

How many deaths would be avoided if males could achieve the same relative survival as females for Melanoma?.

(g) List the avoidable deaths for the oldest age group over all follow-up times. Why are the number of avoidable deaths decreasing as follow-up time increases?

287. **Using `standsurv` for all cause survival and avoidable deaths**

> **Stata addon required!** This exercise requires the Stata user-written command `stpm2` and `standsurv`. See Section 1.2 (page 2) for details and installation instructions.

This question demonstrates the use of `standsurv` to estimate standardized relative and all-cause survival and shows how this can be linked to avoidable deaths.

(a) Load melanoma data, restrict to the 1985-1994 calendar period and merge in the expected mortality rates.

```
. use melanoma, clear
. keep if year8594 == 1
. stset surv_mm, fail(status==1 2) id(id) scale(12) exit(time 60.5)

. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)

. sort _year sex _age
. merge m:1 _year sex _age using popmort,  keep(match master)
```

(b) Generate a dummy for female, generate age splines with 3 df and form interactions between female and the age spline variables

```
. gen female = sex==2
// generate splines
. rcsgen age, gen(agercs) df(3)

// interactions
. forvalues i = 1/3 {
    gen f_agercs‘i’ = female*agercs‘i’
}
```

Fit a model including the effects of sex, age, and their interaction. Allow the effects of age and sex to be time-dependent.

```
. stpm2 female agercs* f_agercs*, scale(h) df(4) ///
        tvc(agercs* female) dftvc(2) bhazard(rate)
```

The coefficients from this model are almost impossible to interpret (in a useful way) individually, but we can use them to make many predictions.

(c) Predict the age standardised relative survival for males and females, using the combined age distribution of males as females as the age standard.

As we have interactions we have to put extra work when we manipulate exposures in our predictions. When we predict for males, we need to make sure the interaction terms are set to zero. When we predict for females we need to make sure the interaction terms are used for all predictions.

```
. range tt 0 5 101
. standsurv, at1(female 0 f_agercs1 0 f_agercs2 0 f_agercs3 0) ///
             at2(female 1 f_agercs1=agercs1 f_agercs2=agercs2 f_agercs3=agercs3) ///
             timevar(tt) ci ///
             atvar(rs_m rs_f)
```

Plot the resulting functions and their 95% confidence intervals.

```
. twoway (line rs_m* tt, lcolor(blue..) lpattern(solid dash dash)) ///
     (line rs_f* tt, lcolor(red..) lpattern(solid dash dash)) ///
    , xtitle("Years from diagnosis") ///
    ytitle("Relative Survival") ///
    ylabel(0.5(0.1)1,angle(h) format(%3.1f)) ///
    legend(order(1 "Males" 4 "Females") ring(0) pos(1) cols(1)) ///
    name(rs, replace)
```

(d) Calculate the survival at 5 years for all subjects where everyone is forced to be male and then female. Take the mean of these estimates. Show these are the same as the estimates at 5 years when using `standsurv`.

```
. gen t5 = 5
. predict rs_ms_m, surv at(female 0 f_agercs1 0 f_agercs2 0 f_agercs3 0) timevar(t5)
. predict rs_ms_f, surv ///
   at(female 1 f_agercs1 = agercs1 f_agercs2 = agercs2 f_agercs3 = . agercs3) timevar(t5)
. tabstat rs_ms_m rs_ms_f
. list rs_m rs_f if tt==5, noobs
```

(e) We will now predict all cause survival by combining relative survival with expected survival, $S_i(t) = S_i^*(t)R_i(t)$. We need to incorporate the population mortality file and various options to estimate the expected survival.

```
standsurv, at1(female 0 f_agercs1 0 f_agercs2 0 f_agercs3 0, atif(female==0)) ///
           at2(female 1 f_agercs1=agercs1 f_agercs2=agercs2 f_agercs3=agercs3, atif(female==1)) ///
           timevar(tt) ci ///
           atvar(acs_m acs_f) ///
           expsurv(using(popmort.dta) ///  popmort file.
             agediag(age)           ///  age at diagnosis
             datediag(dx)           ///  date at diagnosis
             pmage(_age)            ///  age variable in popmort file
             pmyear(_year)          ///  year variable in popmort file
             pmother(sex)           ///  other variables in popmort file
             pmrate(rate)           ///  rate varible in popmort file
             pmmaxyear(2000)        ///  maximum year in popmort file
             at1(sex 1)             ///  variables to match with main at options
             at2(sex 2)             ///  variables to match with main at options
             )
```

As we standardized to the age distribution within each sex, we can compare our modelled based estimate with the Kaplan-Meier estimate. If we see a disagreement then this would indicate that our model was mis-specified.

```
sts gen s_km = s, by(female)
twoway (line acs_m* tt, lcolor(blue..) lpattern(solid dash dash)) ///
       (line acs_f* tt, lcolor(red..) lpattern(solid dash dash)) ///
       (line s_km _t if female == 0, sort lcolor(black) lpattern(shortdash) connect(stairstep)) ///
       (line s_km _t if female == 1, sort lcolor(black) lpattern(shortdash) connect(stairstep)) ///
       , xtitle("Years from diagnosis") ///
       ytitle("All cause Survival") ///
       ylabel(0.5(0.1)1,angle(h) format(%3.1f)) ///
       legend(order(1 "Males" 4 "Females") ring(0) pos(1) cols(1)) ///
       name(acs, replace)
```

Does our model capture the observed marginal all-cause survival?

Give two reasons why comparison of the curves for males and females does not provide a fair comparison in terms of potential differential cancer mortality.

(f) We will now try to quantify the differences between males and females. We first ask the question, "‘how would the relative survival change if males had the same age-specific excess mortality predictions as females?"’

We can restrict the standardisation to males and estimate with and without forcing the excess mortality rates to be those of females. We will also predict for females.

```
standsurv, at1(female 0 f_agercs1 0 f_agercs2 0 f_agercs3 0, atif(female==0)) ///
           at2(female 1 f_agercs1=agercs1 f_agercs2=agercs2 f_agercs3=agercs3, atif(female==0)) ///
           at3(female 1 f_agercs1=agercs1 f_agercs2=agercs2 f_agercs3=agercs3, atif(female==1)) ///
           timevar(tt) ci ///
   atvar(rs2_m_pred rs2_m_withf rs2_f)

twoway (line rs2_m_pred* tt, lcolor(blue..) lpattern(solid dash dash)) ///
       (line rs2_m_withf* tt, lcolor(red..) lpattern(solid dash dash)) ///
       (line rs2_f tt, lcolor(black..) lpattern(dash)) ///
       , xtitle("Years from diagnosis") ///
       ytitle("Relative Survival") ///
       ylabel(0.5(0.1)1,angle(h) format(%3.1f)) ///
       legend(order(1 "Males" 4 "Males (with female RS)" 7 "Females") ring(0) pos(1) cols(1)) ///
       name(rs2, replace)

list rs2_m_pred rs2_m_withf rs2_f if tt==5, noobs
```

Explain why the relative survival of males with females relative survival is not the same as the relative survival of females.

(g) Now we can see what do these difference mean in terms of all-cause survival, i.e. in the real world. We are essentially asking how would the all cause survival change for males, if they had the excess mortality rates of females.

```
standsurv, at1(female 0 f_agercs1 0 f_agercs2 0 f_agercs3 0, atif(female==0)) ///
           at2(female 1 f_agercs1=agercs1 f_agercs2=agercs2 f_agercs3=agercs3, atif(female==0)) ///
           timevar(tt) ci ///
           atvar(acs2_m_pred acs2_m_withf) ///
           expsurv(using(popmort.dta) ///    popmort file.
             agediag(age)            ///    age at diagnosis
             datediag(dx)            ///    date at diagnosis
             pmage(_age)             ///    age variable in popmort file
             pmyear(_year)           ///    year variable in popmort file
             pmother(sex)            ///    other variables in popmort file
             pmrate(rate)            ///    rate varible in popmort file
             pmmaxyear(2000)         ///    maximum year in popmort file
             at1(sex 1)              ///    variables to match with main at options
             at2(sex 1)              ///    variables to match with main at options
             )                       ///
           contrast(difference)      ///
           contrastvar(acs_diff2)

twoway (line acs2_m_pred* tt, lcolor(blue..) lpattern(solid dash dash)) ///
       (line acs2_m_withf* tt, lcolor(red..) lpattern(solid dash dash)) ///
       , xtitle("Years from diagnosis") ///
       ytitle("All cause Survival") ///
       ylabel(0.5(0.1)1,angle(h) format(%3.1f)) ///
       legend(order(1 "Males" 4 "Males (with female RS)") ring(0) pos(1) cols(1)) ///
       name(acs2, replace)

list acs2_m_pred acs2_m_withf  if tt==5, noobs
```

(h) Plot the absolute difference in the survival predicted for males with their own excess mortality rates compared to if the males had the excess mortality rates of the females.

```
twoway (rarea acs_diff2_lci acs_diff2_uci tt, color(red%30)) ///
       (line acs_diff2 tt, lcolor(red..)) ///
       , xtitle("Years from diagnosis") ///
       ytitle("All cause Survival") ///
       ylabel(0.0(0.01)0.1,angle(h) format(%3.2f)) ///
       legend(off) ///
       name(acs2_diff, replace)
```

(i) Rather than just look at the difference we can ask how many fewer deaths would there be if the males had the excess mortality rates of the females. To do this we need to define a population size. One sensible option is to use average number of males diagnosed each year.

Calculate the average number of males diagnosed per year between 1990 and 1994.

```
count if yydx>=1990 & female==0
di `r(N)'/5
```

We will round up to 245 males diagnosed per year We can repeat the previous standsurv command and add the per(245) option. This will just multiply our predictions by 245 and give us the avoidable deaths in a cohort of men diagnosed in a typical calendar year.

```
standsurv, at1(female 0 f_agercs1 0 f_agercs2 0 f_agercs3 0, atif(female==0)) ///
           at2(female 1 f_agercs1=agercs1 f_agercs2=agercs2 f_agercs3=agercs3, atif(female==0)) ///
           timevar(tt) ci ///
           atvar(acs3_m_pred acs3_m_withf) ///
           expsurv(using(popmort.dta) ///   popmort file.
             agediag(age)            ///   age at diagnosis
             datediag(dx)            ///   date at diagnosis
             pmage(_age)             ///   age variable in popmort file
             pmyear(_year)           ///   year variable in popmort file
             pmother(sex)            ///   other variables in popmort file
             pmrate(rate)            ///   rate varible in popmort file
             pmmaxyear(2000)         ///   maximum year in popmort file
             at1(sex 1)              ///   variables to match with main at options
             at2(sex 1)              ///   variables to match with main at options
             )                       ///
           contrast(difference)      ///
           contrastvar(acs_diff3)    ///
           per(245)


twoway (rarea acs_diff3_lci acs_diff3_uci tt, color(red%30)) ///
       (line acs_diff3 tt, lcolor(red..)) ///
       , xtitle("Years from diagnosis") ///
       ytitle("Avoidable Deaths") ///
       ylabel(,angle(h) format(%3.0f)) ///
       legend(off) ///
       name(acs3_diff, replace)
```

Compare this to the estimate in question 282.

284. **Estimating loss in expectation of life**

In this exercise the aim is to estimate the loss in expectation of life for the melanoma cohort as a function of age, year and sex. This can be used to estimate the total number of life years lost for a given cohort of cancer patients. We will also use loss in expectation of life as a way of quantifying the sex difference in melanoma survival, as an alternative to using avoidable deaths (exercise 282).

Loss in expectation of life, together with life expectancy in absence of cancer and life expectancy in presence of cancer can be estimated after fitting a flexible parametric model by using the `lifelost` option of the `predict` postestimation command after using `stpm2` to fit a model. All options used together with `lifelost` are described below:

`mergeby(`*string*`)`	specifies the variables by which the file of general population survival probabilities is sorted.
`diagage(`*name*`)`	specifies the variable containing age at diagnosis. Default is diagage.
`diagyear(`*name*`)`	specifies the variable containing calendar year of diagnosis. Default is diagyear.
`maxage(`*int 99*`)`	specifies the maximum age for which general population survival probabilities are provided in the using file. Probabilities for individuals older than this value are assumed to be the same as for the maximum age. Default is 99.
`attage(`*name*`)`	specifies the variable containing attained age in the popmort file. This variable cannot exist in the patient data file. Default is _age.
`attyear(`*name*`)`	specifies the variable containing attained calendar year in the popmort file. This variable cannot exist in the patient data file. Default is _year.
`survprob(`*name*`)`	specifies the variable containing survival probabilities in the popmort file. This variable cannot exist in the patient data file. Default is prob.
`using(`*string*`)`	specifies the popmort file to be used for expected survival probabilities.
`by(`*string*`)`	specifies stratification variables. Survival probabilities are averaged for each combination of these variables and assumed the same within each combination. Can only be used together with the grpd option.
`maxyear(`*int 2050*`)`	specifies the maximum age for which general population survival probabilities are provided in the using file. Probabilities for years beyond this value are assumed to be the same as for the maximum year. Default is 2050.
`nodes(`*int 50)*	specifies the number of nodes to be used for the numerical integration. Default is 50.
`tinf(`*int 50*`)`	specifies the end year used for the numerical integration. Both observed and expected survival is assumed to be 0 after this point. Default is 50.
`tcond(`*real 0*`)`	specifies the starting year used for the numerical integration. This is used to retrieve conditional estimates. Default is 0.
`grpd`	specifies that average survival probabilities should be used, as opposed to individual probabilities. If this is used together with the by option, the average is calculated within each combination of the specified by variables.
`stub(`*string*`)`	stubname for estimated life expectency in absence and presence of cancer.

(a) Load the melanoma data and `stset` the data for relative survival.

```
. use melanoma, clear
. gen patid = _n
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120.5) id(patid)
```

(b) Fit a flexible parametric model including year, age and sex. Include age and year as continuous variables using splines. Allow all covariates to have a time-dependent effect. Remember to merge on the expected mortality at the exit times.

```
. rcsgen age, df(4) gen(sag) orthog
. rcsgen yydx, df(4) gen(syr) orthog
. gen fem= sex==2

. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort, keep(match master) keepusing(rate)
. drop _age _year _merge

. stpm2 sag1-sag4 syr1-syr4 fem, scale(hazard) df(5) ///
      bhazard(rate) tvc(sag1-sag4 syr1-syr4 fem) dftvc(3)
```

(c) We will now estimate the loss in expectation of life. To save time we don't estimate confidence intervals, although they can be obtained by removing the comments around the `ci` option. (NOTE! Don't attempt to run this with the `ci` option during the lab session. This would take more than an hour, and the only way to stop Stata is to force the program to shut down completely.)

```
. predict ll, lifelost mergeby(_year sex _age) diagage(age) ///
     diagyear(yydx) nodes(40) tinf(80) using(popmort) ///
     stub(surv) maxyear(2000) /*ci*/
```

(d) Create a graph that shows how the loss in expectation of life varies over age, for males diagnosed in 1994.

```
. twoway (line ll age if sex==1 & yydx==1994, sort) , legend(off) ///
    scheme(sj) name(q41_d, replace) ytitle("Years", size(*0.8)) ///
    xtitle("Age at diagnosis", size(*0.8)) xlabel(, labsize(*0.7)) ///
    ylabel(0 5 10 15 20 25 30 35 40 45, labsize(*0.7) angle(0)) ///
    yscale(range(0 45))
```

(e) List the life expectancy and the loss in expectaion of life for someone aged 50, 60, 70 and 80 at diagnosis, both males and females. Also calculate the total number of life years lost among patients diagnosed in 1994.

```
. foreach age in 50 60 70 80 {
    foreach sex in 1 2 {
         list age sex yydx survexp survobs ll if age==‘age’ & ///
         sex==‘sex’ & yydx==1994, constant
    }
   }
. qui summ ll if yydx==1994
. display r(sum)
```

(f) Now estimate the loss in expectation of life if male patients had the same mortality due to melanoma as female patients, but the expected survival of males.

```
. replace fem=1

. predict ll_alt, lifelost mergeby(_year sex _age) diagage(age) ///
     diagyear(yydx) nodes(40) tinf(80) using(popmort) ///
     stub(surv_alt) maxyear(2000) /*ci*/
```

(g) How many life years could potentially be saved if males diagnosed in 1994 had the same survival from melanoma as female patients diagnosed in 1994?

```
. gen lldiff= ll-ll_alt
. summ lldiff if yydx==1994
. display r(sum)

. foreach age in 50 60 70 80 {
     list ll ll_alt lldiff age if sex==1 & age==‘age’ & yydx==1994, constant
   }
```

288. Using standsurv for life expectancy

Stata addon required! This exercise requires the Stata user-written command stpm2 and standsurv. See Section 1.2 (page 2) for details and installation instructions.

This question demonstrates the use of standsurv to estimate life expectancy using the rmst option.

(a) Load melanoma data, restrict to the 1985-1994 calendar period, those aged over 18 and merge in the expected mortality rates.

```
use melanoma, clear
keep if year8594 == 1
stset surv_mm, failure(status=1 2) scale(12) exit(time 120.5) id(id)

gen _age = min(int(age + _t),99) /*merge on expected rates at exittime*/
gen _year = int(yydx + _t)
sort _year sex _age
merge m:1 _year sex _age using popmort, keep(match master) keepusing(rate)
drop _age _year _merge
```

(b) Fit a flexible parametric model including continuous age using restricted cubic splines with 3 knots and and sex. Allow the effect of age and sex to have time-dependent effects.

```
rcsgen age, df(3) gen(agercs)
global ageknots `r(knots)'
gen female = sex==2 // create dummy variable for females

stpm2 agercs* female, scale(hazard) df(5) bhazard(rate) ///
                      tvc(agercs* female) dftvc(3)
```

(c) We will do some predictions for a 70 year old male to show different types of predictions and how to incorporate expected mortality rates using standsurv.

Store the values of the age spline variables for a 70 year old

```
. rcsgen, scalar(70) gen(a) knots(${ageknots})
```

i. Predict relative survival for a 70 year old male up to 10 years and plot.

```
range tt 0 10 101
predict rs70, at(agercs1 `=a1' agercs2 `=a2' agercs3 `=a3' female 0) surv timevar(tt) ci

twoway (line rs70* tt, lcolor(red..) lpattern(solid dash dash)), ///
       ylabel(0(0.1)1, format(%3.1f) angle(h)) ///
       legend(off) ///
       ytitle("Survival") ///
       xtitle("Years from diagnosis") ///
 ylabel(0(0.1)1)
```

ii. The area under the survival curve gives the restricted mean relative survival time. We can approximate this using the integ command.

```
. integ rs70 tt
```

iii. Alternatively we can use the rmst option of stpm2's predict command.

```
. gen t10 = 10 in 1
. predict rs70b in 1, at(agercs1 `=a1' agercs2 `=a2' agercs3 `=a3' female 0) zeros rmst tmax(10)
. list rs70b in 1, noobs
```

Interpret the restricted mean relative survival

(d) Note that we have not extrapolated beyond the range of follow-up, but as this is a parametric model we can extrapolate if we want to. We will extrapolate to 40 years. This is when someone 70 at diagnosis would be 110, so a close to zero probability of being alive.

```
range tt_long 0 40 101
predict rs70_long, at(agercs1 ‘=a1’ agercs2 ‘=a2’ agercs3 ‘=a3’ female 0) surv ///
timevar(tt_long) ci

twoway (line rs70_long* tt_long, lcolor(red..) lpattern(solid dash dash)), ///
       ylabel(0(0.1)1, format(%3.1f) angle(h)) ///
       legend(off) ///
       ytitle("Relative Survival") ///
       xtitle("Years from diagnosis") ///
   ylabel(0(0.1)1)
```

Explain why this is really not a very interesting extrapolation.

(e) We will now combine relative survival with expected survival so we can estimate the all cause probability of death for a 70 year old man. We will switch to `standsurv` for these predictions. We are interested in the prediction of a single individual age 70, so we restict the prediction to the first row and feed in the relevant covariate values using the `at1()` option.

```
gen age70 = 70 in 1
gen dx70 = mdy(1,1,1990)
standsurv if _n==1,                                             ///
           at1(agercs1 ‘=a1’ agercs2 ‘=a2’ agercs3 ‘=a3’ female 0) ///
           timevar(tt) ci             ///
   atvar(s70_all)             ///
           expsurv(using(popmort.dta) ///  popmort file.
             agediag(age70)           ///  age at diagnosis
             datediag(dx70)           ///  date at diagnosis
             pmage(_age)              ///  age variable in popmort file
             pmyear(_year)            ///  year variable in popmort file
             pmother(sex)             ///  other variables in popmort file
             pmrate(rate)             ///  rate varible in popmort file
             pmmaxyear(2000)          ///  maximum year in popmort file
             at1(sex 1)               ///  variables to match with main at options
             expsurvvar(expsurv70))   //   output expected survival

twoway (line s70_all tt) ///
   (line rs70 tt) ///
   (line expsurv70 tt) ///
, ylabel(0(0.1)1, format(%3.1f) angle(h)) ///
legend(order(1 "All cause" 2 "Relative" 3 "Expected") cols(1) ring(0) pos(7)) ///
ytitle("Survival") ///
xtitle("Years from diagnosis")
```

i. Why is the all-cause survival decreasing when the relative survival is flat?

ii. Generate the restricted mean all-cause and expected survival time using `integ` and interpret.

```
integ s70_all tt
integ expsurv70 tt
```

(f) Now repeat, but use `standsurv`. This will perform a more accuarte numerical integration and also give 95% confidence intervals.

```
standsurv if _n==1,  ///
           at1(agercs1 ‘=a1’ agercs2 ‘=a2’ agercs3 ‘=a3’ female 0) ///
           timevar(t10) rmst ci       ///
           atvar(rmst10)              ///
           expsurv(using(popmort.dta) ///  popmort file.
             agediag(age70)           ///  age at diagnosis
             datediag(dx70)           ///  date at diagnosis
             pmage(_age)              ///  age variable in popmort file
             pmyear(_year)            ///  year variable in popmort file
             pmother(sex)             ///  other variables in popmort file
             pmrate(rate)             ///  rate varible in popmort file
             pmmaxyear(2000)          ///  maximum year in popmort file
             at1(sex 1)               ///
             expsurvvar(exp_rmst))

list rmst10* in 1, noobs
list exp_rmst*  in 1, noobs
```

(g) If we use tt_long for our time variable then we will extrapolate both relative and expected survival as a way to extrapolate all-cause survival. In most situations this will work better than direct extrapolation of all cause survival.

```
standsurv if _n==1,  ///
           at1(agercs1 ‘=a1’ agercs2 ‘=a2’ agercs3 ‘=a3’ female 0) ///
           timevar(tt_long)  ci       ///
           atvar(s70_long)                ///
           expsurv(using(popmort.dta) ///  popmort file.
             agediag(age70)           ///  age at diagnosis
             datediag(dx70)           ///  date at diagnosis
             pmage(_age)              ///  age variable in popmort file
             pmyear(_year)            ///  year variable in popmort file
             pmother(sex)             ///  other variables in popmort file
             pmrate(rate)             ///  rate varible in popmort file
             pmmaxyear(2000)          ///  maximum year in popmort file
             at1(sex 1)               ///
             expsurvvar(expsurv70_long))

twoway (line s70_long* tt_long, lcolor(blue..) lpattern(solid dash dash)) ///
       (line rs70_long tt_long) ///
       (line expsurv70_long tt_long) ///
       ,ylabel(0(0.1)1, format(%3.1f) angle(h)) ///
       legend(order(1 "All cause" 4 "Relative" 5 "Expected") cols(1) ring(0) pos(1)) ///
       ytitle("Survival") ///
       xtitle("Years from diagnosis") ///
       xline(10, lpatter(dash))
```

Why can we now integrate the all cause survival curve to obtain an estimate of life expectancy? Perform the integration using integ and interpret.

```
integ s70_long tt_long
integ expsurv70_long tt_long
```

(h) Rather than use integ we can do the estimation using standsurv with the rmst option. Note this is strictly a restricted mean, but we are estimating up to a time point when we know the all-cause survival will be effectively zero.

```
gen t40 = 40 in 1
standsurv if _n==1,  ///
           at1(agercs1 ‘=a1’ agercs2 ‘=a2’ agercs3 ‘=a3’ female 0) ///
           timevar(t40) rmst ci   ///
           atvar(lifeexp70)           ///
           expsurv(using(popmort.dta) ///  popmort file.
             agediag(age70)           ///  age at diagnosis
             datediag(dx70)           ///  date at diagnosis
             pmage(_age)              ///  age variable in popmort file
             pmyear(_year)            ///  year variable in popmort file
             pmother(sex)             ///  other variables in popmort file
             pmrate(rate)             ///  rate varible in popmort file
             pmmaxyear(2000)          ///  maximum year in popmort file
             at1(sex 1)               ///
             expsurvvar(exp_lifeexp70))

list lifeexp70* exp_lifeexp70 in 1, noobs
```

(i) So far we have just estimated for a 70 year old. We can use `standsurv` to obtain standardized survival curves over all (or a subset) of individuals. We need to use the age at diagnosis and date of diagnosis of each individual in the data set.

The youngest person in our analysis is 18 and so extrapolation for just 40 years would only make them 58, so we create a new variable `t100` with 100 years of follow-up. This is longer than necessary for a 80 year old, but is needed as we are now including much younger individuals.

```
gen t100 = 100 in 1
standsurv,  ///
       at1(agercs1 ‘=a1’ agercs2 ‘=a2’ agercs3 ‘=a3’ female 0, atif(female==0)) ///
       timevar(t100) rmst ci   ///
       atvar(lifeexp_stand)           ///
       expsurv(using(popmort.dta)     ///  popmort file.
         agediag(age)                 ///  age at diagnosis
         datediag(dx)                 ///  date at diagnosis
         pmage(_age)                  ///  age variable in popmort file
         pmyear(_year)                ///  year variable in popmort file
         pmother(sex)                 ///  other variables in popmort file
         pmrate(rate)                 ///  rate varible in popmort file
         pmmaxyear(2000)              ///  maximum year in popmort file
         at1(sex 1)                   ///
         expsurvvar(exp_lifeexp_stand))

list lifeexp_stand* exp_lifeexp_stand in 1, noobs
```

Interpret these results.

(j) We need to remember the above is an average over all individuals. This is fine as a summary measure, but hides the fact that life expectency is highly dependent on age. Thus it can be useful to look at life expetency over a range of ages.

We will predict life expectency and the loss in expectation of life for men aged 40, 50, 60, 70 and 80. We just need to extract the values of the age spline variables at these ages.

```
foreach age in 40 50 60 70 80 {
    qui rcsgen, scalar(‘age’) gen(a‘age’_) knots(${ageknots})
    local atopt at1(agercs1 ‘=a‘age’_1’ agercs2 ‘=a‘age’_2’ agercs3 ‘=a‘age’_3’ female 0)
    gen tmpage‘age’ = ‘age’ in 1
    gen tmpdx‘age’ = mdy(1,1,1990) in 1

    standsurv if _n==1, ///
          ‘atopt’                         /// list of at options
          timevar(t100) rmst ci           ///
          atvar(le‘age’)                  ///  new variable names
          expsurv(using(popmort.dta)      ///  popmort file.
            agediag(tmpage‘age’)          ///  age at diagnosis
            datediag(tmpdx‘age’)          ///  date at diagnosis
            pmage(_age)                   ///  age variable in popmort file
            pmyear(_year)                 ///  year variable in popmort file
            pmother(sex)                  ///  other variables in popmort file
            pmrate(rate)                  ///  rate varible in popmort file
            pmmaxyear(2000)               ///  maximum year in popmort file
            at1(sex 1)                    ///
            expsurvvar(exp‘age’))
    drop tmpage‘age’ tmpdx‘age’
}

// Display estimates for selected ages
foreach age in 40 50 60 70 80 {
    di "Age: ‘age’, Cancer = " %5.3f le‘age’[1] ", Expected = " %5.2f exp‘age’
}
```

(k) In question 284 we calculated the average loss in expectation of life if males had the relative survival of females. This can also been done using `standsurv`. We need two `at` options, one where we predict life expectancy for males and one where we predict life expectancy for males, if they had the excess mortality rates of females.

Note we restrict the standardisation to males. As we have no interaction between age and sex, the `at` option is simple. See question 287 on how to code when there are interactions between covariates.
// Also note how we need to specify the `at` suboptions within the `expsurv` option.

```
standsurv, ///
       at1(female 0, atif(female==0)) /// prediction for males
       at2(female 1, atif(female==0)) /// prediction for males (female rs)
       timevar(t40) rmst ci          ///
       atvar(lifeexp_m lifeexp_m_withf) ///
       expsurv(using(popmort.dta)     ///  popmort file.
         agediag(age)                 ///  age at diagnosis
         datediag(dx)                 ///  date at diagnosis
         pmage(_age)                  ///  age variable in popmort file
         pmyear(_year)                ///  year variable in popmort file
         pmother(sex)                 ///  other variables in popmort file
         pmrate(rate)                 ///  rate varible in popmort file
         pmmaxyear(2000)              ///  maximum year in popmort file
         at1(sex 1)                   ///  expected rates for males
         at2(sex 1)                   ///  expected rates for males
         expsurvvar(exp_m1 exp_m2))   ///
        contrast(difference)   ///
contrastvar(lel_stand)

list lifeexp_m lifeexp_m_lci lifeexp_m_uci in 1
list lifeexp_m_withf*  in 1

list lel_stand* in 1
```

3 Solutions

250. **Calculating the crude probability of death from life tables.**

(a) Load the Melanoma data, drop subjects diagnosed 1975-1984 and then and use `strs` to obtain life-tables stratified by age group and sex. Use the `cuminc` option to obtain the crude probabilities of death due to cancer and due to other causes.

```
. stset surv_mm, fail(status==1 2) id(id) scale(12)

                id:  id
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  failure
    t for analysis:  time/12

------------------------------------------------------------------------------
     4744  total observations
        0  exclusions
------------------------------------------------------------------------------
     4744  observations remaining, representing
     4744  subjects
     1404  failures in single-failure-per-subject data
  22108.5  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =  10.95833

. strs using popmort, br(0(1)5) mergeby(_year sex _age) by(agegrp sex) ///
>         save(replace) cuminc list(n d w cp F cp_e2 cr_e2 ci_dc ci_do) f(%7.5f)

         failure _d:  status == 1 2
   analysis time _t:  surv_mm/12
                 id:  id

No late entry detected - p is estimated using the actuarial method
-------------------------------------------------------------------------------------
-> agegrp = 0-44, sex = Male

  +-----------------------------------------------------------------------------------
  | start   end     n    d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_d
  |-----------------------------------------------------------------------------------
  |     0     1   537   25    0   0.95345   0.04655   0.99727   0.95605   0.04389   0.00267
  |     1     2   512   33   43   0.88930   0.11070   0.99437   0.89433   0.10535   0.00535
  |     2     3   436    9   43   0.86999   0.13001   0.99130   0.87762   0.12194   0.00807
  |     3     4   384   18   39   0.82703   0.17297   0.98810   0.83698   0.16216   0.0108
  |     4     5   327    6   34   0.81102   0.18898   0.98473   0.82360   0.17537   0.0136
  +-----------------------------------------------------------------------------------

-------------------------------------------------------------------------------------
-> agegrp = 0-44, sex = Female

  +-----------------------------------------------------------------------------------
  | start   end     n   d    w        cp         F     cp_e2     cr_e2     ci_dc     ci_do
  |-----------------------------------------------------------------------------------
  |     0      1    624    9     0    0.98558    0.01442    0.99911    0.98645    0.01354    0.00088 |
  |     1      2    615    9    52    0.97052    0.02948    0.99816    0.97231    0.02766    0.00182 |
  |     2      3    554    9    56    0.95391    0.04609    0.99712    0.95667    0.04327    0.00282 |
  |     3      4    489    8    51    0.93745    0.06255    0.99599    0.94122    0.05867    0.00389 |
  |     4      5    430    8    68    0.91851    0.08149    0.99477    0.92334    0.07647    0.00503 |
  +-------------------------------------------------------------------------------------------------+

-----------------------------------------------------------------------------------------------------
-> agegrp = 45-59, sex = Male

  +---------------------------------------------------------------------------------------------------+
  | start    end      n    d     w          cp           F      cp_e2      cr_e2      ci_dc      ci_do |
  |---------------------------------------------------------------------------------------------------|
  |     0      1    752   51     0     0.93218     0.06782    0.99094    0.94070    0.05903    0.00879 |
  |     1      2    701   38    72     0.87891     0.12109    0.98140    0.89557    0.10353    0.01755 |
  |     2      3    591   38    64     0.81917     0.18083    0.97111    0.84354    0.15433    0.02650 |
  |     3      4    489   17    61     0.78879     0.21121    0.96025    0.82145    0.17566    0.03554 |
  |     4      5    411   16    53     0.75597     0.24403    0.94866    0.79688    0.19912    0.04491 |
  +---------------------------------------------------------------------------------------------------+

-----------------------------------------------------------------------------------------------------
-> agegrp = 45-59, sex = Female

  +---------------------------------------------------------------------------------------------------+
  | start    end      n    d     w          cp           F      cp_e2      cr_e2      ci_dc      ci_do |
  |---------------------------------------------------------------------------------------------------|
  |     0      1    612   21     0     0.96569     0.03431    0.99661    0.96897    0.03098    0.00333 |
  |     1      2    591   23    61     0.92606     0.07394    0.99298    0.93261    0.06715    0.00679 |
  |     2      3    507   16    64     0.89487     0.10513    0.98906    0.90477    0.09474    0.01039 |
  |     3      4    427   11    62     0.87001     0.12999    0.98482    0.88341    0.11581    0.01418 |
  |     4      5    354    5    49     0.85681     0.14319    0.98034    0.87399    0.12508    0.01812 |
  +---------------------------------------------------------------------------------------------------+

-----------------------------------------------------------------------------------------------------
-> agegrp = 60-74, sex = Male

  +---------------------------------------------------------------------------------------------------+
  | start    end      n    d     w          cp           F      cp_e2      cr_e2      ci_dc      ci_do |
  |---------------------------------------------------------------------------------------------------|
  |     0      1    709   61     0     0.91396     0.08604    0.96735    0.94481    0.05429    0.03175 |
  |     1      2    648   67    75     0.81366     0.18634    0.93361    0.87152    0.12395    0.06239 |
  |     2      3    506   37    63     0.75021     0.24979    0.89794    0.83548    0.15695    0.09283 |
  |     3      4    406   39    55     0.67291     0.32709    0.86090    0.78164    0.20430    0.12279 |
  |     4      5    312   27    51     0.60950     0.39050    0.82214    0.74135    0.23821    0.15230 |
  +---------------------------------------------------------------------------------------------------+

-----------------------------------------------------------------------------------------------------
-> agegrp = 60-74, sex = Female

  +---------------------------------------------------------------------------------------------------+
  | start    end      n    d     w          cp           F      cp_e2      cr_e2      ci_dc      ci_do |
  |---------------------------------------------------------------------------------------------------|
  |     0      1    661   41     0     0.93797     0.06203    0.98381    0.95340    0.04622    0.01581 |
  |     1      2    620   47    60     0.86325     0.13675    0.96623    0.89343    0.10470    0.03205 |
  |     2      3    513   31    62     0.80773     0.19227    0.94730    0.85267    0.14369    0.04857 |
  |     3     4   420    22    52   0.76263   0.23737   0.92670   0.82295   0.17154   0.0658
  |     4     5   346    18    48   0.72000   0.28000   0.90473   0.79582   0.19638   0.0836
  +------------------------------------------------------------------------------------------
```

```
--------------------------------------------------------------------------------------------
-> agegrp = 75+, sex = Male

  +-----------------------------------------------------------------------------------------+
  | start   end     n    d    w         cp          F      cp_e2      cr_e2      ci_dc      ci_do |
  |-----------------------------------------------------------------------------------------|
  |     0     1   337   67    0    0.80119    0.19881    0.88853    0.90170    0.09282    0.10599 |
  |     1     2   270   61   37    0.60686    0.39314    0.78562    0.77247    0.20100    0.19214 |
  |     2     3   172   33   17    0.48438    0.51562    0.68883    0.70319    0.25207    0.26355 |
  |     3     4   122   19   19    0.40257    0.59743    0.59992    0.67104    0.27279    0.32464 |
  |     4     5    84   11   12    0.34580    0.65420    0.52181    0.66269    0.27747    0.37673 |
  +-----------------------------------------------------------------------------------------+

--------------------------------------------------------------------------------------------
-> agegrp = 75+, sex = Female

  +-----------------------------------------------------------------------------------------+
  | start   end     n    d    w         cp          F      cp_e2      cr_e2      ci_dc      ci_do |
  |-----------------------------------------------------------------------------------------|
  |     0     1   512   68    0    0.86719    0.13281    0.91552    0.94721    0.05056    0.08225 |
  |     1     2   444   75   47    0.71252    0.28748    0.83184    0.85655    0.12977    0.15772 |
  |     2     3   322   50   32    0.59609    0.40391    0.75041    0.79436    0.17897    0.22494 |
  |     3     4   240   39   27    0.49345    0.50655    0.67530    0.73072    0.22433    0.28221 |
  |     4     5   174   23   24    0.42340    0.57660    0.60436    0.70057    0.24363    0.33298 |
  +-----------------------------------------------------------------------------------------+
```

(b) How is the probability of death due to all causes, F, calculated?

This is just 1 - the survival function , i.e. `1-cp`.

(c) Why is the crude probability of death due to cancer, `ci_dc` similar to the all-cause probability of death for subjects aged 0-44?

```
. use grouped, clear
(Collapsed (or grouped) survival data)

. list agegrp start end sex F ci_dc if agegrp == 0 & sex == 1, noobs
  +-------------------------------------------------+
  | agegrp   start   end    sex         F     ci_dc |
  |-------------------------------------------------|
  |   0-44       0     1   Male   0.04655   0.04389 |
  |   0-44       1     2   Male   0.11070   0.10535 |
  |   0-44       2     3   Male   0.13001   0.12194 |
  |   0-44       3     4   Male   0.17297   0.16216 |
  |   0-44       4     5   Male   0.18898   0.17537 |
  +-------------------------------------------------+
```

They are similar as there is low probability that subjects of this age will die from other causes. Thus, if they die it is highly likely to be due to cancer.

(d) For both males and females aged 60-74 what is the probability of death due to all causes at 5 years post diagnosis? What two variables can be added together to give the probability of death due to all-causes?}

```
. list  end agegrp sex F ci_dc ci_do if  agegrp == 2 & end == 5
     +-----------------------------------------------------+
     | end   agegrp      sex         F     ci_dc     ci_do |
     |-----------------------------------------------------|
 25. |   5    60-74     Male   0.39050   0.23821   0.15230 |
 30. |   5    60-74   Female   0.28000   0.19638   0.08362 |
     +----------------------------------------------------+

. gen F2 = ci_dc + ci_do
. list  end agegrp sex F ci_dc ci_do F2 if  agegrp == 2 & end == 5
     +-----------------------------------------------------------------+
     | end   agegrp      sex         F     ci_dc     ci_do          F2 |
     |-----------------------------------------------------------------|
 25. |   5    60-74     Male   0.39050   0.23821   0.15230    .3905036 |
 30. |   5    60-74   Female   0.28000   0.19638   0.08362    .2800009 |
     +-----------------------------------------------------------------+
```

The probability of death due to all causes is 0.39 for males and 0.28 for females. With crude mortality we partition the all-cause probability of death into that due to cancer and that due to other cause. Thus `F = ci_dc + ci_do`.

(e) What proportion of the all-cause deaths at 5 years post diagnosis are due to cancer and due to other causes for males? Compare these figures for the different age groups.

```
. gen prob_c = ci_dc / F
. gen prob_o = ci_do / F
. list  end agegrp sex F ci_dc ci_do prob_c prob_o ///
>        if  end == 5 & sex == 1, noobs

  +-----------------------------------------------------------------------+
  | end   agegrp   sex         F     ci_dc     ci_do     prob_c    prob_o |
  |-----------------------------------------------------------------------|
  |   5     0-44   Male   0.18898   0.17537   0.01361     .92796  .0720402 |
  |   5    45-59   Male   0.24403   0.19912   0.04491   .8159498  .1840501 |
  |   5    60-74   Male   0.39050   0.23821   0.15230   .6100003  .3899997 |
  |   5      75+   Male   0.65420   0.27747   0.37673   .4241378  .5758622 |
  +-----------------------------------------------------------------------+
```

In the youngest age group 93% of the deaths are associated with a diagnosis of cancer at 5 years poist diagnosis. In the oldest agegroup the figure is 42%. This is due to increased probability of dying from other causes in the oldest age group.

(f) The age groups are fairly wide, explain how you would expect the crude probability of death due to cancer to differ between a 60 and 74 year old, even if the relative survival was identical.

Since the probability of death due to other cause is higher for a 74 year old than for a 60 year old then if relative survival was identical we would expect the actual probability of death due to cancer to be lower for someone aged 74 than a 60 year old.

(g) Plot the net probability of death, the crude probability of death due to cancer and the overall probability of death for males by age group. Try to understand the relationship between these various measures.

```
. gen net = 1- cr_e2

. twoway  (line F net ci_dc end if sex == 1, sort ), by(agegrp) ///
>                legend(order(1 "Overall" 2 "Net" 3 "Crude") cols(3)) ///
>                ylabel(0(0.1)0.6, angle(h) format(%3.1f)) ///
>                ytitle("Probability of Death")
```

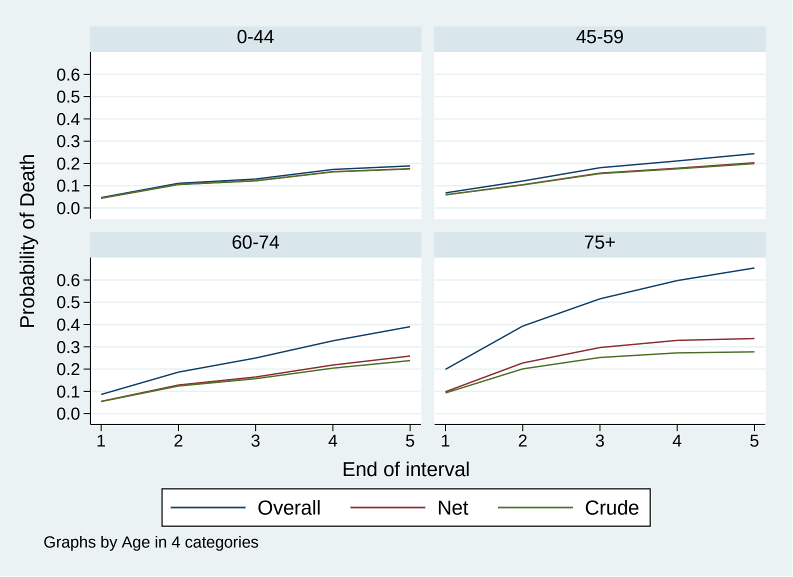

Figure 1: Melanoma Data. All cause, Net and Crude Probability of Death due to cancer.

Very little difference between the estimates in youngest age group. Increasing separation as age increases due to increased contribution of deaths due to other causes.

251. **Probability of death in a competing risks framework (relative survival model)**

In exercise 250 we explored how one could estimate crude probabilities of death based on life table estimates of relative survival making use of the strs implementation of the approach proposed by Cronin and Feuer (2000) [7]. Lambert *et al.* (2010) [8] subsequently showed how the estimates can be obtained after fitting a relative survival model, namely a flexible parametric models for relative survival, which use restricted cubic splines for the baseline cumulative excess hazard and for any time-dependent effects. Although the two approaches estimate the same quantity, the life table approach provides estimates for grouped data so we get an estimated probability for an age group rather than an estimate for a specific age as can be obtained in the model-based approach.

(a) Load the Melanoma data and merge in the background mortality rates as in question **??**. Fit a flexible parametric relative survival model including age group with time-dependent effects.

```
. tab agegrp, gen(agegrp)
. stpm2 agegrp2-agegrp4, scale(hazard) bhazard(rate) df(5) ///
      tvc(agegrp2-agegrp4) dftvc(2)
```

Calculate the estimated net probability of death (1 - relative survival) and plot the four curves on a single graph. Interpret the plot.

(b) Use the standsurv command to estimate the crude probability of death. Note that standsurv will predict for individual covariate patterns and for specific ages at diagnosis. Perform the predictions for males aged 40, 55, 70 and 80 diagnosed in 1985. As we are only making predictions for one individual, we need to create a variable with age at diagnosis and date at diagnosis for the healthy individual to match to. This is used as the prediction for the expected survival. We also define a user-defined mata function calc_allcause to calculate the all-cause failure function as a sum of the two at() options. The prediction for a 40 year old (the first age group) can be obtained using,

```
. mata function calc_allcause(at) return(at[1]+at[2])

. range temptime2 0 5 101
. gen aged = .
. gen dated = mdy(1,1,1985) in 1
. replace aged = 40 in 1

. standsurv if _n==1,  at1(sex 1 agegrp2 0 agegrp3 0 agegrp4 0) ///
>  verbose timevar(temptime2) ///
>  atvar(crprob1) crudeprob stub2(cancer other) ///
>  expsurv(using("Z:\cansurv\data\popmort.dta") ///
>          datediag(dated)                ///
>          agediag(aged)               ///
>          pmrate(rate)                       ///
>          pmage(_age)                                ///
>          pmyear(_year)              ///
>          pmother(sex)                       ///
>          pmmaxyear(1985)                 ///
>          at1(sex 1))                                        ///
>  userfunction(calc_allcause) ///
>  userfunctionvar(allcause1) transform(none)
```

Plot the estimated crude probability of death due to cancer for each of the selected ages on the same graph. Contrast these with the estimated net probability of death from part (a).

(c) Generate a similar plot but for the crude probability of death due to other causes.

(d) A useful way of presenting crude probabilities is through stacked graphs.

  i. Generate the stacked graphs for each of the selected ages. Use the solution Do file for help.

  ii. Now overlay the net probability of death. Does it better illustrate the contrast described in (b)?

(e) Advanced: Now fit a model using splines for the effect age with the spline terms allowed to be time-dependent.

  i. Calculate the crude probabilities of death and compare these to the model where age is categorized.

  ii. Now calculate crude probabilities of death at individual ages from 40 to 90 years old at 5 years since diagnosis - plot these over age. See do file for help. Hint: you will need to do a loop over 50 `standsurv` predictions.

260. **Parametric cure models**

(a) `_t` contains the time in years from diagnosis. The `strsmix` command requires the expected mortality rate at the event time. The first `gen` command calculates the age at the event (or censoring) time (up to a maximum age of 99). The second `gen` command calculates the calender year at the event time. The third `gen` command converts the expected survival probability into the expected mortality rate.

(b) Fitting this model gives

```
. strsmix if year8594==0, dist(weibull) link(identity) bhazard(rate)

                                                  Number of obs   =       6477
                                                  Wald chi2(0)    =          .
Log likelihood =  -9988.719                       Prob > chi2     =          .

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
       _cons |   .4151695   .0081152    51.16   0.000      .399264     .431075
-------------+----------------------------------------------------------------
ln_lambda    |
       _cons |  -.1694096   .0257529    -6.58   0.000    -.2198843   -.1189348
-------------+----------------------------------------------------------------
ln_gamma     |
       _cons |  -.1783506   .0166044   -10.74   0.000    -.2108946   -.1458066
------------------------------------------------------------------------------
```

i. The cure fraction is 0.415 (i.e. 41.5%).

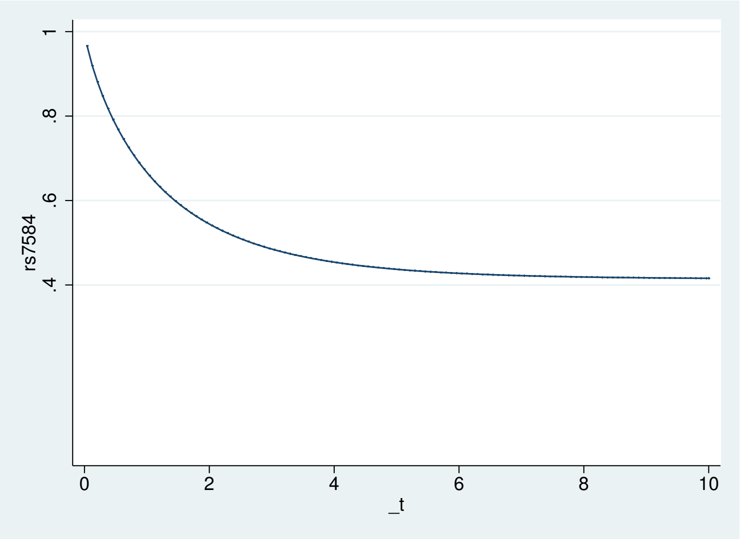

Figure 2: Relative survival in 1975-1984 for cancer of the colon

ii. Yes the relative survival curves reaches a plateau at the cure fraction. Note that if this did not appear to be the case then the cure fraction estimate would be based on extrapolation beyond the range of follow-up in the data.

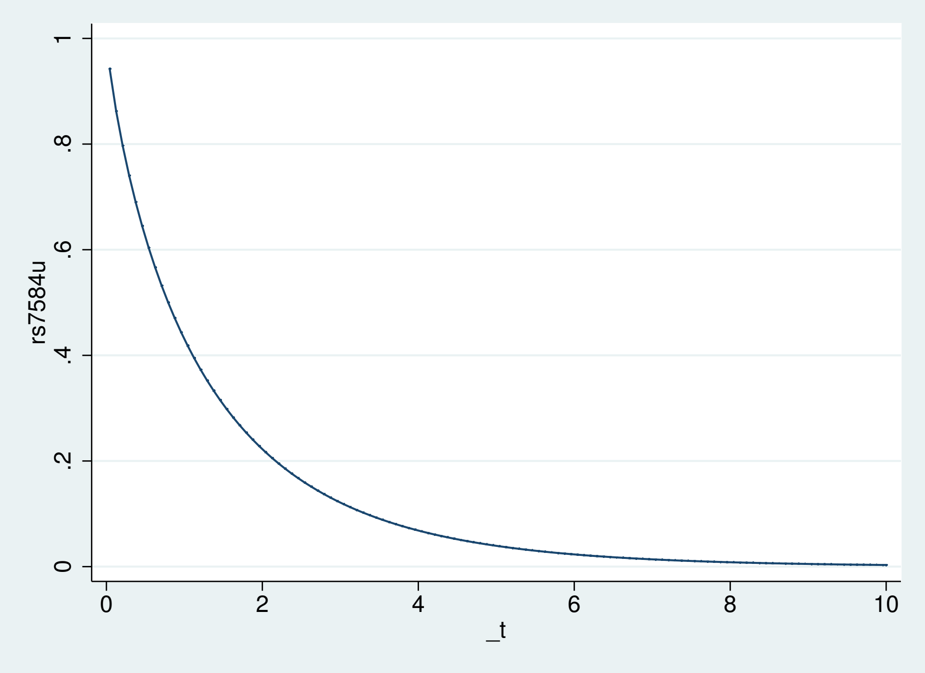

Figure 3: Relative survival for the 'uncured' in 1975-1984 for cancer of the colon

iii. Approximately 80% of the 'uncured' have died after 2 years.

iv. Median survival for the 'uncured' is approximately 0.8 years

(c) Now fitting to those diagnosed 1985-1994.

```
. strsmix if year8594==1, dist(weibull) link(identity) bhazard(rate)

                                                  Number of obs   =       9087
                                                  Wald chi2(0)    =          .
Log likelihood = -11339.861                       Prob > chi2     =          .

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
       _cons |     .46044   .0087593    52.57   0.000     .4432721    .4776078
-------------+----------------------------------------------------------------
ln_lambda    |
       _cons |  -.2648208   .0292473    -9.05   0.000    -.3221445   -.2074972
-------------+----------------------------------------------------------------
ln_gamma     |
       _cons |  -.2101828   .0163283   -12.87   0.000    -.2421857   -.1781799
------------------------------------------------------------------------------
```

i. The cure fraction is now 0.459 (i.e 45.9%) - a difference of 4.5%.

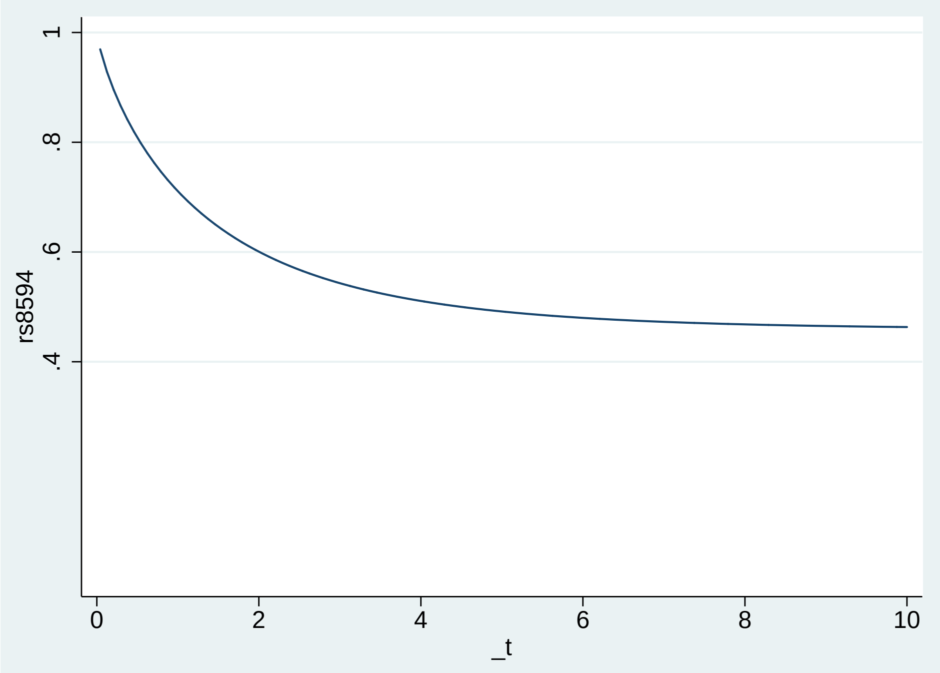

Figure 4: Relative survival in 1985-1984 for cancer of the colon

ii. Yes, the relative survival cure reaches a plateau.

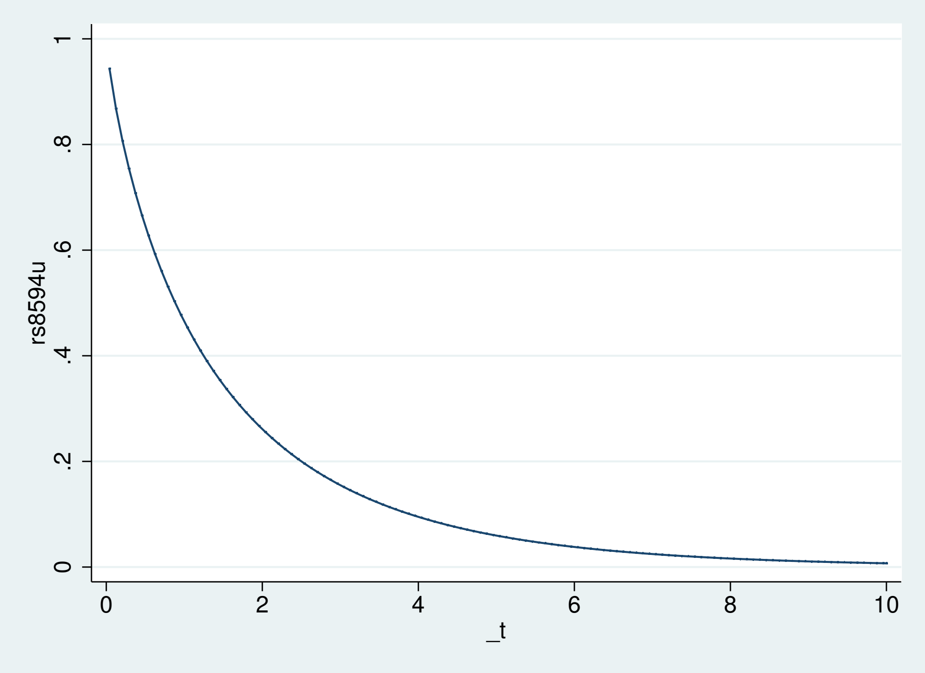

Figure 5: Relative survival for the 'uncured' in 1975-1984 for cancer of the colon

iii. At two years about 75% of the 'uncured' have died after 2 years. A reduction of about 5% in absolute terms.

iv. The median survival of the 'uncured' is about 0.9 years, a slight improvement.

(d) Including `year8594` as a covariate gives

```
. strsmix year8594, dist(weibull) link(identity) bhazard(rate)


                                                   Number of obs   =      15564
                                                   Wald chi2(1)    =      38.51
Log likelihood =  -21332.05                        Prob > chi2     =     0.0000
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
    year8594 |   .0618817   .0099714     6.21   0.000      .042338    .0814254
       _cons |   .4090526   .0078184    52.32   0.000     .3937288    .4243765
-------------+----------------------------------------------------------------
ln_lambda    |
       _cons |  -.2110754   .0191294   -11.03   0.000    -.2485684   -.1735825
-------------+----------------------------------------------------------------
ln_gamma     |
       _cons |  -.1925967   .0115469   -16.68   0.000    -.2152282   -.1699652
------------------------------------------------------------------------------
```

i. The estimated difference in the cure fraction is 0.062 (i.e. 6.2%). This is larger than the difference observed in b(i) and c(i).

ii. The assumption is that the survival distribution of the 'uncured' is the same in the two periods. This is because λ and γ do not vary by our covariate (`year8594`).

   Allowing both λ and γ to vary by `year8594` gives

```
. strsmix year8594, dist(weibull) link(identity) bhazard(rate) ///
k1(year8594) k2(year8594)


                                                   Number of obs   =      15564
                                                   Wald chi2(1)    =      14.37
Log likelihood =  -21328.58                        Prob > chi2     =     0.0001

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
    year8594 |   .0452705   .0119408     3.79   0.000     .0218671     .068674
       _cons |   .4151695   .0081152    51.16   0.000      .399264     .431075
-------------+----------------------------------------------------------------
ln_lambda    |
    year8594 |  -.0954111   .0389694    -2.45   0.014    -.1717897   -.0190325
       _cons |  -.1694096   .0257529    -6.58   0.000    -.2198843   -.1189348
-------------+----------------------------------------------------------------
ln_gamma     |
    year8594 |  -.0318322   .0232878    -1.37   0.172    -.0774754     .013811
       _cons |  -.1783506   .0166044   -10.74   0.000    -.2108946   -.1458066
------------------------------------------------------------------------------
```

iii. The difference in the cure fraction is 0.045 (i.e. 4.5%). This gives the same as we observed when fitting two separate models, as this is essentially what we are doing by including `year8594` for all 3 parameters. If the distribution of the 'uncured' is not modelled appropriately then biased estimates of the cure fraction may be obtained.

iv. Using a Wald test gives

```
. test [ln_lambda][year8594] [ln_gamma][year8594], mtest
```

```
 ( 1)  [ln_lambda]year8594 = 0
 ( 2)  [ln_gamma]year8594 = 0
---------------------------------------
       |        chi2     df       p
-------+-------------------------------
  (1)  |        6.00      1     0.0143 #
  (2)  |        1.83      1     0.1761 #
-------+-------------------------------
  all  |        6.84      2     0.0328
---------------------------------------
                  # unadjusted p-values
```

There is evidence that the survival distribution of the 'uncured' differs between the two time periods.

(e) This model can be fitted using the xi prefix command.

```
. tab agegrp, gen(cage)
 strsmix year8594 cage1 cage2 cage3 cage4, dist(weibull) link(logit) ///
        bhazard(rate) k1(year8594 cage1 cage2 cage3 cage4) ///
        k2(year8594 cage1 cage2 cage3 cage4) eform

                                                  Number of obs   =      15564
                                                   Wald chi2(4)    =      28.29
Log likelihood = -21088.807                        Prob > chi2     =     0.0000

------------------------------------------------------------------------------
          _t |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pi           |
    year8594 |   1.231615   .0573756     4.47   0.000     1.124142    1.349363
       cage2 |    .903997   .0879128    -1.04   0.299     .7471167    1.093819
       cage3 |   .7988555    .072884    -2.46   0.014     .6680492    .9552742
       cage4 |    .869293    .080983    -1.50   0.133     .7242167    1.043431
       _cons |    .891236   .0760408    -1.35   0.177     .7539937    1.053459
-------------+----------------------------------------------------------------
ln_lambda    |
    year8594 |  -.1118244   .0392174    -2.85   0.004     -.188689   -.0349597
       cage2 |   .0856077    .084418     1.01   0.311    -.0798484    .2510639
       cage3 |   .2501009   .0791222     3.16   0.002     .0950243    .4051775
       cage4 |    1.00063   .0845808    11.83   0.000     .8348543    1.166405
       _cons |  -.5465794   .0750655    -7.28   0.000    -.6937052   -.3994537
-------------+----------------------------------------------------------------
ln_gamma     |
    year8594 |  -.0241314   .0224827    -1.07   0.283    -.0681968     .019934
       cage2 |  -.0614646    .056022    -1.10   0.273    -.1712656    .0483365
       cage3 |  -.1322088   .0518933    -2.55   0.011    -.2339179   -.0304997
       cage4 |  -.1330111   .0527858    -2.52   0.012    -.2364693   -.0295528
       _cons |  -.0000647   .0498729    -0.00   0.999    -.0978138    .0976845
------------------------------------------------------------------------------
```

i. The parameter estimates for the cure fraction are now odds ratios. Thus the odds of cure are 23% higher in 1985-1994 when compared to 1975-1984. For age group 0-44 is the reference category. The odds of cure are 10% lower in the 45-59 age group, 21% lower in the 60-74 age group and 14% lower in the 75+ age group. Only the 60-84 age group is significant at the 5% level. The needs to be a degree of caution here as the Weibull cure models tends to not fit well to the oldest age group and more complex models may be necessary.

ii. The predicted median survival for the ‘uncured’ is obtained using

```
. predict med, centile
. bysort agegrp year8594: gen flag = (_n==1)
. list agegrp year8594 med if flag==1, noobs

  +---------------------------------------+
  | agegrp          year8594          med |
  |---------------------------------------|
  |   0-44   Diagnosed 75-84     1.197311 |
  |   0-44   Diagnosed 85-94    1.3485631 |
  |  45-59   Diagnosed 75-84     1.105672 |
  |  45-59   Diagnosed 85-94    1.2519877 |
  |  60-74   Diagnosed 75-84    .92317295 |
  |---------------------------------------|
  |  60-74   Diagnosed 85-94    1.0500786 |
  |    75+   Diagnosed 75-84    .39166079 |
  |    75+   Diagnosed 85-94    .43631407 |
  +---------------------------------------+
```

This table shows how median survival increases with time period in each age group. In addition median survival for the ‘uncured’ decreases with age.

261. **Estimating cure models using flexible parametric survival models**

(a)

```
. stpm2 year8594, df(6) bhazard(rate) scale(hazard) cure

Iteration 0:   log likelihood = -21851.481
Iteration 1:   log likelihood = -21147.216
Iteration 2:   log likelihood = -21095.674
Iteration 3:   log likelihood = -21095.385
Iteration 4:   log likelihood = -21095.385

Log likelihood = -21095.385                       Number of obs   =      15564

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |  -.1556103    .025088    -6.20   0.000    -.2047819   -.1064388
       _rcs1 |   .9889082   .0117887    83.89   0.000     .9658028    1.012014
       _rcs2 |   .0353623    .006665     5.31   0.000      .022299    .0484255
       _rcs3 |   .0684074   .0045871    14.91   0.000     .0594168     .077398
       _rcs4 |   .0530653   .0039162    13.55   0.000     .0453896     .060741
       _rcs5 |   .0410339   .0032154    12.76   0.000     .0347319    .0473359
       _rcs6 |  (omitted)
       _cons |  -.1110995   .0197347    -5.63   0.000    -.1497788   -.0724201
------------------------------------------------------------------------------
```

i. The coefficient -.1556103 is the log-hazard ratio (HR = 0.86) comparing the second period to the first.

ii. The cure proportion for the first period is $\exp(-\exp(-.1110995)) = .40866901$, and for the second period $\exp(-\exp(-.1110995 - .1556103)) = .4649175$.

iii.

```
. predict cure1, cure

. list cure1 if year8594==0, constant

  +---------+
  |   cure1 |
  |---------|
  | .408669 |
  +---------+
  (no variables vary in 6477 observations)

. list cure1 if year8594==1, constant

  +-----------+
  |     cure1 |
  |-----------|
  | .46491749 |
  +-----------+
  (no variables vary in 9087 observations)
```

iv. The estimated difference in the cure fraction is 0.056 (i.e. 5.6%) compared to 0.062 (i.e. 6.2%) in exercise 260.

v. The predicted median survival times are similar in the two groups, but not the same. The flexible parametric cure model is a special case of a non-mixture model. Non-mixture cure models use both the estimated cure proportions and the specified distribution function to estimate the survival function of uncured, which will lead to different survival even when no time-dependent effects are modelled.

```
. predict med1, centile(50) uncured

. list med1 if year8594==0, constant

  +-----------+
  |      med1 |
  |-----------|
  | .75329265 |
  +-----------+
  (no variables vary in 6477 observations)

. list med1 if year8594==1, constant

  +-----------+
  |      med1 |
  |-----------|
  | .80035703 |
  +-----------+
  (no variables vary in 9087 observations)
```

(b)

```
. stpm2 year8594, df(6) tvc(year8594) dftvc(4) bhazard(rate) scale(hazard) cure

Iteration 0:   log likelihood = -21848.799
Iteration 1:   log likelihood = -21144.251
Iteration 2:   log likelihood = -21092.538
Iteration 3:   log likelihood = -21092.239
Iteration 4:   log likelihood = -21092.239

Log likelihood = -21092.239                     Number of obs   =      15564

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
    year8594 |  -.1492647   .0269617    -5.54   0.000    -.2021086   -.0964208
       _rcs1 |   1.006746   .0177333    56.77   0.000     .9719896    1.041503
       _rcs2 |   .0447082   .0094731     4.72   0.000     .0261413    .0632751
       _rcs3 |   .0692846   .0065112    10.64   0.000     .0565229    .0820462
       _rcs4 |   .0493157   .0057847     8.53   0.000     .0379779    .0606535
       _rcs5 |   .0384908   .0038595     9.97   0.000     .0309262    .0460553
       _rcs6 |  (omitted)
_rcs_y~85941 |  -.0329169   .0238804    -1.38   0.168    -.0797216    .0138878
_rcs_y~85942 |  -.0137549   .0135084    -1.02   0.309    -.0402309    .0127211
_rcs_y~85943 |   .0100166   .0086015     1.16   0.244    -.0068419    .0268752
_rcs_y~85944 |  (omitted)
       _cons |  -.1131936   .0202657    -5.59   0.000    -.1529136   -.0734736
------------------------------------------------------------------------------
```

i. The coefficient is no longer interpreted as the log-hazard ratio since the hazard ratio is varying over time.

ii. The cure proportion for the first period is $\exp(-\exp(-.1131936)) = 0.40943474$, and for the second period $\exp(-\exp(-.1131936 - .1492647)) = 0.46340289$.

iii.

```
. predict cure2, cure
. list cure2 if year8594==0, constant
  +-----------+
  |     cure2 |
  |-----------|
  | .40943473 |
  +-----------+
  (no variables vary in 6477 observations)

. list cure2 if year8594==1, constant
  +-----------+
  |     cure2 |
  |-----------|
  | .46340288 |
  +-----------+
  (no variables vary in 9087 observations)
```

iv. The estimated difference in the cure fraction is 0.054 (i.e. 5.4%), very similar to the result in a.

v. The difference in the predicted median survival times between the two groups is larger than in a, since we are now allowing more flexibility into the estimation.

```
. predict med2, centile(50) uncured

. list med2 if year8594==0, constant
  +----------+
  |     med2 |
  |----------|
  | .7406603 |
  +----------+
  (no variables vary in 6477 observations)

. list med2 if year8594==1, constant
  +-----------+
  |      med2 |
  |-----------|
  | .81717336 |
  +-----------+
  (no variables vary in 9087 observations)
```

(c) The flexible parametric cure model forces the cumulative excess hazard to be constant after the last knot, and therefore the relative survival is forced to reach a plateau. The assumption of cure should always be checked in a model that does not assume cure or by looking at empirical life table estimates.

```
. predict surv, survival
. predict survunc, survival uncured
. forvalues j=0/1 {
        twoway (line surv _t if year8594==`j', sort) ///
          (line survunc _t if year8594==`j', sort), ///
          legend(label(1 "Survival overall") ///
          label(2 "Survival for uncured")) name(period`j', replace)
  }
```

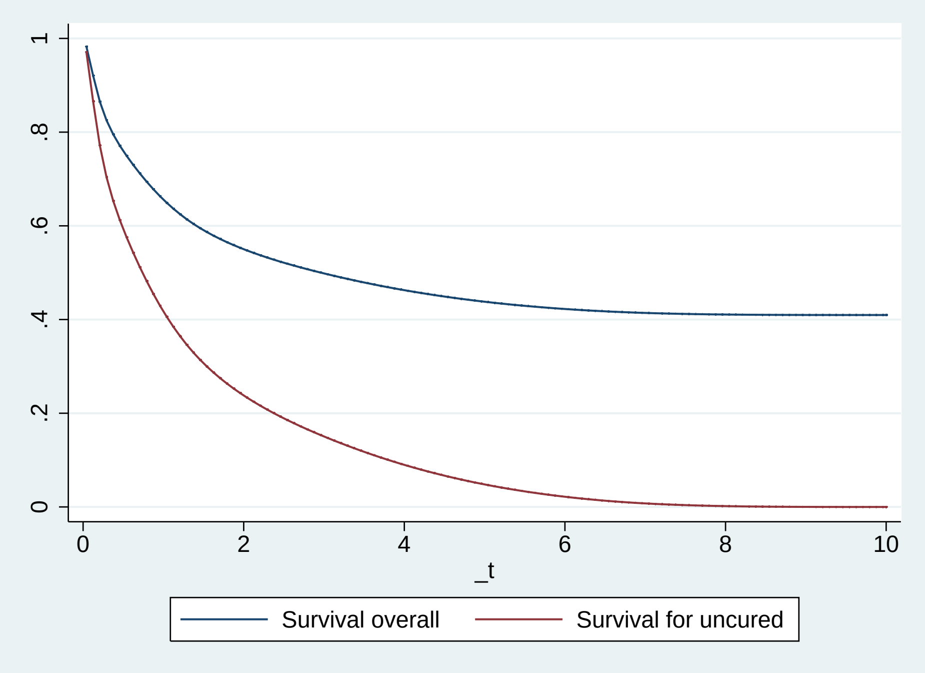

Figure 6: Relative survival overall and for the ‘uncured’ in 1975-1984 for cancer of the colon

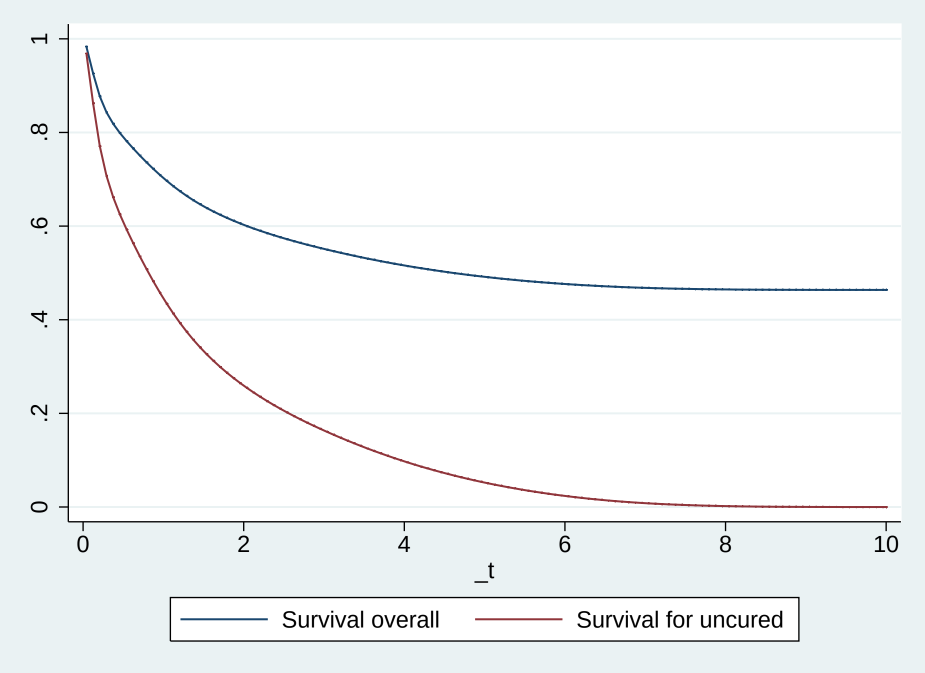

Figure 7: Relative survival overall and for the ‘uncured’ in 1985-1994 for cancer of the colon

282. **Calculating excess and 'avoidable' deaths from life tables.**

(a) Load the Melanoma data, drop subjects diagnosed 1975-1984.

(b) What is the difference in five-year relative survival between males and females in each age group?

```
. list agegrp sex cr_e2 if end == 5, noobs sepby(agegrp)
  +--------------------------+
  | agegrp      sex    cr_e2 |
  |--------------------------|
  |   0-44     Male   0.8236 |
  |   0-44   Female   0.9233 |
  |--------------------------|
  |  45-59     Male   0.7969 |
  |  45-59   Female   0.8740 |
  |--------------------------|
  |  60-74     Male   0.7413 |
  |  60-74   Female   0.7958 |
  |--------------------------|
  |    75+     Male   0.6627 |
  |    75+   Female   0.7006 |
  +--------------------------+
```

Five year relative survival is lower for males in all age groups.

(c) Reshape the data.

```
. bysort sex (agegrp start): gen j = _n
. gen sexlab =cond(sex==1,"_m","_f")
. drop sex
. reshape wide start end n cp cp_e2 cr_e2 agegrp, i(j) j(sexlab) string
(note: j = _f _m)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       40   ->      20
Number of variables                   9   ->      15
j variable (2 values)            sexlab   ->   (dropped)
xij variables:
                                  start   ->   start_f start_m
                                    end   ->   end_f end_m
                                      n   ->   n_f n_m
                                     cp   ->   cp_f cp_m
                                  cp_e2   ->   cp_e2_f cp_e2_m
                                  cr_e2   ->   cr_e2_f cr_e2_m
                                 agegrp   ->   agegrp_f agegrp_m
-----------------------------------------------------------------------------

. rename agegrp_m agegrp
. rename start_m start
. rename end_m end
. drop agegrp_f start_f end_f
```

(d) For males, calculate the expected number of all-cause deaths, Nd_m, the expected number of deaths if the study population were free of cancer, NExp_d_m and the excess deaths associated with a diagnosis of cancer, ED_m.

```
. bys agegrp: gen Nrisk_m = n_m[1]/10

. gen p_dead_m = 1 - cp_e2_m * cr_e2_m
. gen Nd_m = Nrisk_m*p_dead_m
. gen NExp_d_m = Nrisk_m*(1-cp_e2_m)
. gen ED_m = Nd_m - NExp_d_m

. format Nd_m NExp_d_m ED_m %4.1f
. list agegrp Nrisk_m p_dead_m Nd_m NExp_d_m ED_m if end==5, noobs
  +-----------------------------------------------------+
  | agegrp   Nrisk_m   p_dead_m   Nd_m   NExp_d_m   ED_m |
  |-----------------------------------------------------|
  |   0-44      53.7   .1889797   10.1        0.8    9.3 |
  |  45-59      75.2   .2440302   18.4        3.9   14.5 |
  |  60-74      70.9   .3905036   27.7       12.6   15.1 |
  |    75+      33.7   .6542017   22.0       16.1    5.9 |
  +-----------------------------------------------------+

. table agegrp if end == 5, c(sum Nd_m sum NExp_d_m sum ED_m) row format(%4.1f)
-------------------------------------------------------
_m agegrp |     sum(Nd_m)  sum(NExp_d_m)      sum(ED_m)
----------+--------------------------------------------
     0-44 |          10.1            0.8            9.3
    45-59 |          18.4            3.9           14.5
    60-74 |          27.7           12.6           15.1
      75+ |          22.0           16.1            5.9
          |
    Total |          78.2           33.4           44.8
-------------------------------------------------------
```

i. We would expect to see 10, 18, 28 and 22 all cause deaths in the (ascending) age groups.

ii. This is given by the excess deaths, ED_m. In ascending age groups there are 9, 14, 15, and 6 excess deaths at 5 years post diagnosis when compared to a similar cancer free population. This is for a typical cohort diagnosed in one calendar year.

iii. There are 45 excess deaths when compared to the general population.

(e) Repeat calculations for females.

```
. bys agegrp: gen Nrisk_f = n_f[1]/10

. gen p_dead_f = 1 - cp_e2_f * cr_e2_f
. gen Nd_f = Nrisk_f*p_dead_f
. gen NExp_d_f = Nrisk_f*(1-cp_e2_f)
. gen ED_f = Nd_f - NExp_d_f

. format Nd_f NExp_d_f ED_f %4.1f
. list agegrp Nrisk_f p_dead_f Nd_f NExp_d_f ED_f if end==5, noobs
  +-----------------------------------------------------+
  | agegrp   Nrisk_f   p_dead_f   Nd_f   NExp_d_f   ED_f |
  |-----------------------------------------------------|
  |   0-44      62.4   .0814915    5.1        0.3    4.8 |
  |  45-59      61.2   .1431934    8.8        1.2    7.6 |
  |  60-74      66.1   .2800009   18.5        6.3   12.2 |
  |    75+      51.2   .5766043   29.5       20.3    9.3 |
  +---------------------------------------------------+

. table agegrp if end == 5, c(sum Nd_f sum NExp_d_f sum ED_f) row format(%4.1f)
------------------------------------------------------
_m agegrp |     sum(Nd_f)  sum(NExp_d_f)      sum(ED_f)
----------+-------------------------------------------
     0-44 |           5.1            0.3            4.8
    45-59 |           8.8            1.2            7.6
    60-74 |          18.5            6.3           12.2
      75+ |          29.5           20.3            9.3
          |
    Total |          61.9           28.1           33.8
------------------------------------------------------
```

In terms of the total number of all cause deaths, females have fewer at all ages except the 70+ group. This is because they are more females diagnosed in this group 51 vs 34, so even though females have lower relative survival they have more deaths due to a number of women in the oldest age groups being diagnosed. This leads to there being more excess deaths in this age group for women when compared to men. As a whole there are more excess deaths in men.

(f) How many deaths would be ‘avoided’ if males could achieve the same relative survival as females for Melanoma?

```
. gen Nd_m_f = Nrisk_m*(1 - cp_e2_m * cr_e2_f)
. gen AD_m = Nd_m - Nd_m_f

. format Nd_m_f AD_m %4.1f
. list agegrp Nrisk_m p_dead_m Nd_m NExp_d_m ED_m Nd_m_f AD_m if end==5, noobs
  +-------------------------------------------------------------------------+
  | agegrp   Nrisk_m   p_dead_m   Nd_m   NExp_d_m   ED_m   Nd_m_f    AD_m |
  |-------------------------------------------------------------------------|
  |   0-44      53.7   .1889797   10.1        0.8    9.3      4.9     5.3 |
  |  45-59      75.2   .2440302   18.4        3.9   14.5     12.9     5.5 |
  |  60-74      70.9   .3905036   27.7       12.6   15.1     24.5     3.2 |
  |    75+      33.7   .6542017   22.0       16.1    5.9     21.4     0.7 |
  +-------------------------------------------------------------------------+
```

There would be about 15 deaths ‘avoided’. The youngest two age groups contribute most to the avoidable deaths.

(g) List the avoidable deaths for the oldest age group over all follow-up times. Why are the number of avoidable deaths decreasing as follow-up time increases?

```
. list agegrp end AD_m if agegrp==3
     +---------------------+
     | agegrp   end   AD_m |
     |---------------------|
 16. |    75+     1    1.4 |
 17. |    75+     2    2.2 |
 18. |    75+     3    2.1 |
 19. |    75+     4    1.2 |
 20. |    75+     5    0.7 |
     +---------------------+
```

This is because we can not avoid deaths for ever. Remember that we are looking at all cause deaths. If we had unlimited follow-up we would avoid no deaths at all. In the oldest age group we can actually see that we are just postponing deaths.

284. Estimating loss in expectation of life

(a) Load the Melanoma data and `stset` the data for relative survival.

```
. use melanoma, clear
(Skin melanoma, diagnosed 1975-94, follow-up to 1995)
. gen patid = _n
. stset surv_mm, failure(status=1 2) scale(12) exit(time 120.5) id(patid)

                id:  patid
     failure event:  status == 1 2
obs. time interval:  (surv_mm[_n-1], surv_mm]
 exit on or before:  time 120.5
    t for analysis:  time/12
------------------------------------------------------------------------------
     7775  total observations
        0  exclusions
------------------------------------------------------------------------------
     7775  observations remaining, representing
     7775  subjects
     2777  failures in single-failure-per-subject data
 43384.63  total analysis time at risk and under observation
                                              at risk from t =         0
                                   earliest observed entry t =         0
                                        last observed exit t =  10.04167
```

(b) Fit a flexible parametric model including year, age and sex. Include age and year as continuous variables using splines. Allow all covariates to have a time-dependent effect. Remember to merge on the expected mortality at the exit times.

```
. rcsgen age, df(4) gen(sag) orthog
Variables sag1 to sag4 were created

. rcsgen yydx, df(4) gen(syr) orthog
Variables syr1 to syr4 were created

. gen fem = sex==2
. gen _age = min(int(age + _t),99)
. gen _year = int(yydx + _t)
. sort _year sex _age
. merge m:1 _year sex _age using popmort, keep(match master) keepusing(rate)

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                             7,775  (_merge==3)
    -----------------------------------------
. drop _age _year _merge

. stpm2 sag1-sag4 syr1-syr4 fem, scale(hazard) df(5) bhazard(rate) ///
>                 tvc(sag1-sag4 syr1-syr4 fem) dftvc(3)
```

```
Log likelihood = -8444.5801                      Number of obs   =       7775
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
xb           |
        sag1 |   .3486966   .0355765     9.80   0.000     .2789678    .4184253
        sag2 |  -.0382469   .0368393    -1.04   0.299    -.1104506    .0339568
        sag3 |  -.0826459   .0352677    -2.34   0.019    -.1517692   -.0135225
        sag4 |  -.0171397   .0333635    -0.51   0.607     -.082531    .0482516
        syr1 |  -.0064674   .1187121    -0.05   0.957    -.2391387     .226204
        syr2 |  -.2522286   .1030806    -2.45   0.014    -.4542629   -.0501944
        syr3 |  -.1413523   .0858927    -1.65   0.100     -.309699    .0269943
        syr4 |  -.1155111   .0700542    -1.65   0.099    -.2528149    .0217927
         fem |  -.5220707   .0604833    -8.63   0.000    -.6406158   -.4035256
       _rcs1 |   .9474817   .0781558    12.12   0.000     .7942992    1.100664
       _rcs2 |   .1927113    .054332     3.55   0.000     .0862225    .2992001
       _rcs3 |   .0568751   .0304669     1.87   0.062    -.0028389    .1165892
       _rcs4 |   .0032183    .014089     0.23   0.819    -.0243957    .0308323
       _rcs5 |   .0063443   .0052562     1.21   0.227    -.0039577    .0166462
  _rcs_sag11 |   .0101007   .0305454     0.33   0.741    -.0497673    .0699687
  _rcs_sag12 |   .0327253    .026622     1.23   0.219    -.0194529    .0849034
  _rcs_sag13 |   .0204141   .0135927     1.50   0.133     -.006227    .0470553
  _rcs_sag21 |  -.0382793   .0312975    -1.22   0.221    -.0996212    .0230626
  _rcs_sag22 |  -.0024951   .0278919    -0.09   0.929    -.0571622    .0521719
  _rcs_sag23 |   .0015633   .0139492     0.11   0.911    -.0257767    .0289032
  _rcs_sag31 |  -.0148982   .0288652    -0.52   0.606     -.071473    .0416766
  _rcs_sag32 |   .0178845    .025579     0.70   0.484    -.0322494    .0680183
  _rcs_sag33 |   .0007745   .0129807     0.06   0.952    -.0246672    .0262163
  _rcs_sag41 |  -.0217533   .0278767    -0.78   0.435    -.0763907    .0328841
  _rcs_sag42 |   .0036575   .0247048     0.15   0.882    -.0447631    .0520781
  _rcs_sag43 |  -.0002257   .0126263    -0.02   0.986    -.0249727    .0245214
  _rcs_syr11 |   .1082871   .0951937     1.14   0.255    -.0782891    .2948633
  _rcs_syr12 |  -.0912392   .0569474    -1.60   0.109    -.2028541    .0203757
  _rcs_syr13 |  -.0598222   .0368902    -1.62   0.105    -.1321258    .0124813
  _rcs_syr21 |  -.1088465   .0811995    -1.34   0.180    -.2679946    .0503015
  _rcs_syr22 |   .0769735   .0481734     1.60   0.110    -.0174446    .1713916
  _rcs_syr23 |   .0206394    .030727     0.67   0.502    -.0395845    .0808632
  _rcs_syr31 |  -.1046798   .0660342    -1.59   0.113    -.2341045    .0247448
  _rcs_syr32 |   .0236841   .0431332     0.55   0.583    -.0608553    .1082236
  _rcs_syr33 |   .0266358   .0243036     1.10   0.273    -.0209984      .07427
  _rcs_syr41 |  -.0203372   .0520008    -0.39   0.696    -.1222569    .0815826
  _rcs_syr42 |   .0493604   .0349461     1.41   0.158    -.0191328    .1178536
  _rcs_syr43 |   .0196377   .0188815     1.04   0.298    -.0173694    .0566448
   _rcs_fem1 |  -.0019995   .0503392    -0.04   0.968    -.1006625    .0966635
   _rcs_fem2 |  -.0844331   .0450417    -1.87   0.061    -.1727131     .003847
   _rcs_fem3 |  -.0203553   .0212678    -0.96   0.339    -.0620393    .0213288
       _cons |  -1.378518   .0959111   -14.37   0.000      -1.5665   -1.190535
------------------------------------------------------------------------------
```

(c) We will now estimate the loss in expectation of life. To save time we don't estimate confidence intervals, although they can be obtained by removing the comments around the `ci` option.

```
. predict ll, lifelost mergeby(_year sex _age) diagage(age) diagyear(yydx) nodes(40) tinf(8
>                                 using(popmort) stub(surv) maxyear(2000) /*ci*/
```

(d) Create a graph that shows how the loss in expectation of life varies over age, for males diagnosed in 1994.

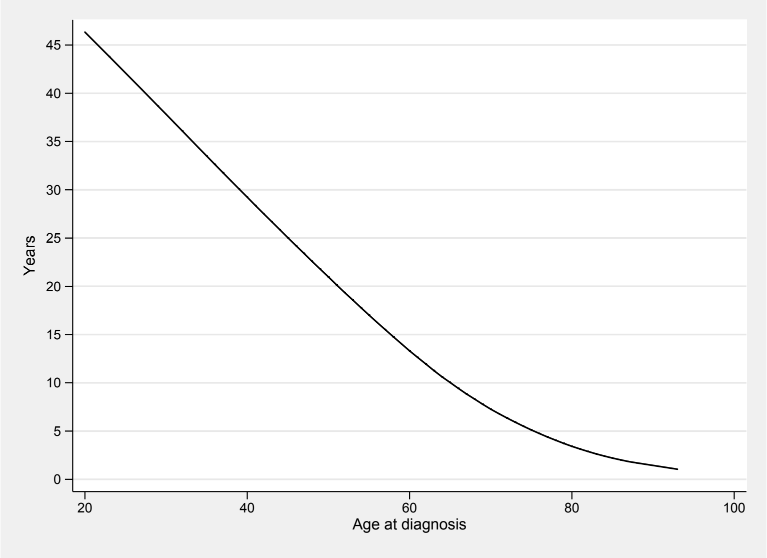

Figure 8: Melanoma Data. Loss in expectation of life

Figure 8 shows the loss in expectation of life for males diagnosed with melanoma in 1994.

(e) List the life expectancy and the loss in expectation of life for someone aged 50, 60, 70 and 80 at diagnosis, both males and females. Also calculate the total number of life years lost among patients diagnosed in 1994.

```
. foreach age in 50 60 70 80 {
  2.        foreach sex in 1 2 {
  3.                list age sex yydx survexp survobs ll if age==‘age’ & sex==‘sex’ & yydx==
  4.        }
  5. }

  +------------------------------------------------------+
  | age    sex   yydx    survexp     survobs         ll |
  |------------------------------------------------------|
  |  50   Male   1994   26.63637   5.6614445   20.97493 |
  +------------------------------------------------------+
  (no variables vary in 5 observations)

  +--------------------------------------------------------+
  | age      sex   yydx    survexp     survobs         ll |
  |--------------------------------------------------------|
  |  50   Female   1994   32.36633   7.2172614   25.14907 |
  +--------------------------------------------------------+
  (no variables vary in 3 observations)

  +------------------------------------------------------+
  | age    sex   yydx    survexp     survobs         ll |
  |------------------------------------------------------|
  |  60   Male   1994   18.49159   5.1773682   13.31423 |
  +------------------------------------------------------+
(no variables vary in 8 observations)
```

```
  +-------------------------------------------------------+
  | age      sex   yydx    survexp     survobs         ll |
  |-------------------------------------------------------|
  |  60   Female   1994   23.30669   6.8167728   16.48991 |
  +-------------------------------------------------------+
  (no variables vary in 8 observations)

  +----------------------------------------------------+
  | age    sex   yydx    survexp     survobs        ll |
  |----------------------------------------------------|
  |  70   Male   1994   11.53323   4.2612695   7.27196 |
  +----------------------------------------------------+
  (no variables vary in 4 observations)

  +------------------------------------------------------+
  | age      sex   yydx   survexp     survobs         ll |
  |------------------------------------------------------|
  |  70   Female   1994   14.8622   5.8554623   9.006738 |
  +------------------------------------------------------+
  (no variables vary in 9 observations)

  +-----------------------------------------------------+
  | age    sex   yydx    survexp     survobs         ll |
  |-----------------------------------------------------|
  |  80   Male   1994   6.431057   3.0075134   3.423544 |
  +-----------------------------------------------------+
  (no variables vary in 3 observations)

  +-------------------------------------------------------+
  | age      sex   yydx    survexp     survobs         ll |
  |-------------------------------------------------------|
  |  80   Female   1994   8.000338   4.1340081   3.866329 |
  +-------------------------------------------------------+
  (no variables vary in 3 observations)

. qui summ ll if yydx==1994
. display r(sum)
8767.1307
```

The total number of life years lost among patients diagnosed with melanoma in Finland in 1994 is 8767.

(f) Now estimate the loss in expectation of life if male patients had the same mortality due to melanoma as female patients, but the expected survival of males.

```
. replace fem=1
(3680 real changes made)

. predict ll_alt, lifelost mergeby(_year sex _age) diagage(age) diagyear(yydx) nodes(40) tinf
>                                  using(popmort) stub(surv_alt) maxyear(2000) /*ci*/
```

(g) How many life years could potentially be saved if males diagnosed in 1994 had the same survival from melanoma as female patients diagnosed in 1994?

```
. gen lldiff= ll-ll_alt
. summ lldiff if yydx==1994

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      lldiff |       518    .6344759    .6386128          0   1.554199

. display r(sum)
328.6585

. foreach age in 50 60 70 80 {
  2.   list ll ll_alt lldiff age if sex==1 & age==‘age’ & yydx==1994, constant
  3. }

  +-------------------------------------+
  |       ll     ll_alt    lldiff   age |
  |-------------------------------------|
  | 20.97493   19.56192   1.41301    50 |
  +-------------------------------------+
  (no variables vary in 5 observations)

  +------------------------------------+
  |       ll     ll_alt   lldiff   age |
  |------------------------------------|
  | 13.31423   11.99303   1.3212    60 |
  +------------------------------------+
  (no variables vary in 8 observations)

  +-------------------------------------+
  |      ll     ll_alt     lldiff   age |
  |-------------------------------------|
  | 7.27196   6.200533   1.071427    70 |
  +-------------------------------------+
  (no variables vary in 4 observations)

  +--------------------------------------+
  |       ll     ll_alt     lldiff   age |
  |--------------------------------------|
  | 3.423544   2.734462   .6890819    80 |
  +--------------------------------------+
```

If males diagnosed in 1994 had the same relative survival as females diagnosed in 1994, the total number of life years lost would reduce by 328 years. For a man aged 50 at diagnosis the potential gain in life expectancy is 1.4 years (1.3, 1.1 and 0.7 years for males aged 60, 70 and 80 years at diagnosis, respectively).

References

[1] Dickman PW, Coviello E. Estimating and modelling relative survival. *The Stata Journal* 2015;**15**:186–215.

[2] Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. *The Stata Journal* 2009;**9**:265–290.

[3] Andersson TML, Lambert PC. Fitting and modeling cure in population-based cancer studies within the framework of flexible parametric survival models. *The Stata Journal* 2012; **12**:623–628.

[4] Lambert PC. Modeling of the cure fraction in survival studies. *The Stata Journal* 2007; **7**:351–375.

[5] Hinchliffe SR, Lambert PC. Extending the flexible parametric survival model for competing risks. *The Stata Journal* 2013;**13**:344–355.

[6] Corazziari I, Quinn M, Capocaccia R. Standard cancer patient population for age standardising survival ratios. *Eur J Cancer* 2004;**40**:2307–2316.

[7] Cronin KA, Feuer EJ. Cumulative cause-specific mortality for cancer patients in the presence of other causes: a crude analogue of relative survival. *Statistics in Medicine* 2000; **19**:1729–1740.

[8] Lambert PC, Dickman PW, Nelson CP, Royston P. Estimating the crude probability of death due to cancer and other causes using relative survival models. *Statistics in Medicine* 2010;**29**:885 – 895.

[9] Lambert PC, Dickman PW, Österlund P, Andersson TML, Sankila R, Glimelius B. 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* 2007; **121**:2052–2059.

[10] Lambert PC, Thompson JR, Weston CL, Dickman PW. Estimating and modeling the cure fraction in population-based cancer survival analysis. *Biostatistics* 2007;**8**:576–594.

[11] Belot A, Ndiaye A, Luque-Fernandez MA, Kipourou DK, Maringe C, Rubio FJ, Rachet B. Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clinical Epidemiology* 2019;**11**:53–65.