Estimating and modelling relative survival using SAS

Paul Dickman

October 18, 2004

Contents

1 Quick start

This document describes SAS code (version 7 or higher) for estimating and modelling relative survival. A sample data set containing information on colon carcinoma diagnosed in Finland is provided. All required data and SAS files can be downloaded in a ZIP archive from:

`http://www.pauldickman.com/rsmodel/sas_colon.zip`

If the files are extracted to the directory `c:\rsmodel\sas_colon\` then the code should run without requiring alteration. The code provided will reproduce the estimates reported in Table I of the paper by Dickman *et al.* [1]. Two input data files are provided; `colon.sas7bdat` contains the cancer patient data and `popmort.sas7bdat` contains data on expected probabilities of death for the Finnish general population.

Running the SAS code in `survival.sas` will produce life table estimates of relative survival stratified by sex, age, and calendar period of diagnosis. In addition, two output data sets are created (one containing grouped data and one containing individual patient data) which are used as input data sets for modelling. The SAS code in `models.sas` estimates a relative survival regression model using several different approaches (described in Dickman *et al.* [1]).

In general, two data files are required in order to estimate relative survival; a file containing individual-level data on the patients (see section 3.2) and a file containing expected probabilities of death for a comparable general population (see section 3.3). To estimate and model relative survival using other data it will almost certainly be necessary to modify the code to allow for differences in, for example, variable names and formats. Constructing your data set in a similar fashion to the example data sets (e.g., using identical variable names) is probably the easiest way to get started.

2 Estimating relative survival

The relative survival ratio (RSR) is estimated using life table methods — the cumulative RSR is estimated at discrete points in the follow-up by taking the product of interval-specific estimates over sub-intervals of the follow-up. For example, to estimate cumulative 10-year survival we might estimate conditional survival (interval-specific survival) for each of 10 annual intervals and then multiply the interval-specific estimates to obtain the cumulative estimates. Following is an example of a resulting life table:

```
Life table estimates of patient survival
Males aged 0-44 diagnosed with colon carcinoma in Finland 1975-84

                  Interval-            Interval-            Interval-
                  specific  Cumulative specific  Cumulative specific  Cumulative
                  observed   observed  expected   expected  relative   relative
Int.     N  D  W  survival   survival  survival   survival  survival   survival

0 - 1   75  4  0  0.94667    0.94667   0.99697    0.99697   0.94954    0.94954
1 - 2   71  8  0  0.88732    0.84000   0.99682    0.99381   0.89015    0.84524
2 - 3   63  1  1  0.98400    0.82656   0.99649    0.99032   0.98747    0.83464
3 - 4   61  3  0  0.95082    0.78591   0.99625    0.98660   0.95440    0.79658
4 - 5   58  3  0  0.94828    0.74526   0.99601    0.98266   0.95208    0.75841
5 - 6   55  2  0  0.96364    0.71816   0.99562    0.97836   0.96787    0.73404
6 - 7   53  0  0  1.00000    0.71816   0.99532    0.97378   1.00470    0.73749
7 - 8   53  0  0  1.00000    0.71816   0.99491    0.96882   1.00512    0.74127
8 - 9   53  1  0  0.98113    0.70461   0.99453    0.96352   0.98653    0.73128
9 - 10  52  2  0  0.96154    0.67751   0.99418    0.95792   0.96717    0.70727
```

3 Overview of the approach to estimating relative survival in SAS

My approach is to 'split' the observation time for each individual into multiple observations, one for each band of follow-up time (where the bands correspond to life table intervals). That is, we start with one observation for each individual but then 'split' this to obtain one observation for each life table interval. This is identical to the approach commonly used for analysing epidemiological cohort studies where each individual's time at risk is classified according to factors such as time since entry, attained age, period, and cohort. To split the data, we use the SAS macro `lexis` (written by Bendix Carstensen [2]).

The general approach is as follows:

1. Use the `lexis` macro to split the observation time for each individual into multiple observations, one for each band of follow-up time (i.e., one for each life table interval).
2. Ensure each record has the correct value for attained age and attained calendar period.
3. Merge, from an external file (the so-called 'popmort' file), expected probabilities of surviving the interval.
4. For each observation, create indicator variables for death and censoring.
5. Create life tables for each desired combination of covariates by collapsing over relevant records. That is, all observation that refer to, for example, the third life table interval for a class of patients are 'collapsed' to form a single observation. The number of observations collapsed to form this single observation is equal to the number first at risk in the interval. Summing the indicator variables for dead and censored gives the total number of deaths and censorings in the interval. The average of the subject-specific expected survival probabilities gives the Ederer II estimate of expected survival for the interval.
6. For each life table interval, calculate interval-specific and cumulative observed, expected, and relative survival with corresponding standard errors and confidence intervals.
7. Print the life tables (using PROC PRINT).

Models for relative survival can be estimated based on either individual (subject-specific) or grouped/collapsed (i.e., life table) data. As such, a copy of the data is saved both before collapsing (the default data set name is `individ`) and after collapsing (the default data set name is `grouped`).

This approach to estimating survival has several advantages:

- It provides output data sets to which a range of approaches to model fitting can be applied.
- Using these same output data sets, it is possible to produce tables or graphs of estimates.
- The algorithm and the code are transparent and can easily be modified/adapted by the user.
- Both traditional cohort-based estimates as well as period estimates of patient survival can be obtained.

3.1 Contents of the ZIP archive

The following files are contained in the ZIP archive `sas_colon.zip`. If these files are extracted to the directory `c:\rsmodel\sas_colon\` then the code should run without requiring alteration. The code provided will reproduce the estimates reported in Table I of the paper by Dickman *et al.* [1].

`colon.sas7bdat` Patient data file containing a single observation for each individual diagnosed with colon carcinoma in Finland 1975-1994 with follow-up to the end of 1995 (SAS version 7/8 format).

`popmort.sas7bdat` Expected survival probabilities for the Finnish general population stratified by age, period, and sex (SAS version 7/8 format).

`survival.sas` SAS code for estimating patient survival using traditional cohort analysis and creating two output files which can be used for modelling.

`survival_period.sas` SAS code for estimating patient survival using period analysis and creating two output files which can be used for modelling.

`models.sas` SAS code for estimating various relative survival models using the output from `survival.sas` or `survival_period.sas`.

`formats.sas` SAS code (PROC FORMAT) for creating the format catalogue.

`formats.sas7bcat` SAS format catalogue created by `formats.sas`.

`lexis.sas` Lexis macro for splitting person-time.

`table_of_survival_estimates.sas` SAS code for producing a table of 1, 5, 10, and 20-year relative survival estimates (to be run after `survival.sas`).

`graph_of_survival_estimates.sas` SAS code for producing a graph of relative survival estimates (to be run after `survival.sas`).

3.2 The patient data file (`colon.sas7bdat`)

The patient data file (`colon.sas7bdat`) in the distribution contains individual-level data for 15,564 patients diagnosed with colon carcinoma in Finland 1975-1994 with follow-up to the end of 1995. In the examples presented here we study only the 6,274 patients with localised tumours.

```
Variable   Type  Len  Format     Label
---------------------------------------------------------------------
AGE        Num     4             Age at diagnosis
DX         Num     8  DATE.      Date of diagnosis
EXIT       Num     8  DATE.      Date of exit
MMDX       Num     4             Month of diagnosis
SEX        Num     4  SEX.       Sex
STAGE      Num     4  STAGE.     Clinical stage at diagnosis
STATUS     Num     4  STATUS.    Vital status at last date of contact
SUBSITE    Num     4  COLONSUB.  Anatomical subsite of tumour
SURV_MM    Num     4             Survival time in completed months
SURV_YY    Num     4             Survival time in completed years
YEAR8594   Num     4             Indicator for year of dx 1985-94
YYDX       Num     4             Year of diagnosis
```

In general, a file containing information on individuals diagnosed with cancer is required and must contain, at a minimum, the following information:

- Survival time (time at risk). In the examples from Finland survival time has been calculated in advance although it is possible to specify the date of diagnosis and date of exit (see section 4.0.2 on page 8).
- Indicator for vital status (dead/alive). Information on cause of death is not required.
- Variables upon which expected survival depends – typically age, sex, and period (as in the Finnish example) but can also include, for example, race, region/country of residence, or social class.

3.3 The population mortality file (popmort.sas7bdat)

The population mortality file (popmort.sas7bdat) contains expected survival probabilities (PROB) for the Finnish general population stratified by age, calendar year, and sex for the years 1951 to 2000. Although the popmort file contains expected probabilities for each year, they were calculated by the central statical office (Statistics Finland) for 5-year intervals so are identical, for example, for all years from 1951–1955.

An extract of the first 20 observations in the file is shown below.

```
SEX    _YEAR    _AGE      PROB

 1      1951      0     0.96429
 1      1951      1     0.99639
 1      1951      2     0.99783
 1      1951      3     0.99842
 1      1951      4     0.99882
 1      1951      5     0.99893
 1      1951      6     0.99913
 1      1951      7     0.99905
 1      1951      8     0.99920
 1      1951      9     0.99931
 1      1951     10     0.99940
 1      1951     11     0.99939
 1      1951     12     0.99920
 1      1951     13     0.99925
 1      1951     14     0.99914
 1      1951     15     0.99913
 1      1951     16     0.99897
 1      1951     17     0.99882
 1      1951     18     0.99835
 1      1951     19     0.99836
```

In general, this file should be stratified by all variables upon which expected survival depends – typically age, sex, and period (as in the Finnish example) but can also include, for example, race, region/country of residence, or social class.

Survival is often estimated for subgroups defined by year of diagnosis or age at diagnosis. When estimating expected survival it is the age and year at time of follow-up (rather than at the time of diagnosis) that are important. That is, our data file will contain variables for both age at diagnosis and attained age (age at the time of follow-ups). I have adopted the convention of prefixing variable names with an underscore when they are updated with follow-up, for example, the variable age contains age at diagnosis and _age contains attained age.

4 Illustration of the code for estimating survival (`survival.sas`)

The approach is implemented as SAS code, rather than as a SAS macro, in order to make it more transparent and easier to customize. The main parameters are, however, defined as macro variables at the top of the file. For example, we first define the two input data files and specify filenames for the two output data files.

```
/* Population mortality file */
%let popmort=colon.popmort ;

/* Patient data file */
%let patdata=colon.colon ;

/* Output data file containing individual records */
%let individ=colon.individ ;

/* Output data file containing collapsed data */
%let grouped=colon.grouped ;
```

The macro variable `vars` stores the variables over which the life tables are stratified. The example below results in a lifetable being estimated for each combination of sex, yydx (year of diagnosis), and age (at diagnosis). A single life table is estimated for every combination of each value of these variables. Formats can be used to group metric variables into categories. For example, the variable `age` contains age in completed years and a format is specified to group this into categories.

```
%let vars = sex yydx age;
%let formats = sex sex. age age. yydx yydx. ;
```

The formats are defined in `formats.sas` and stored in the permanent format library `colon.formats`. The `FMTSEARCH` option is used (near the top of the code) to define the search path for formats.

The next step is a data step where housekeeping on the patient data file is performed. These tasks can, of course, be performed before running `survival.sas` but I have found it practical to include them at this stage. For example, we exclude observations not eligible for the analysis (patients with stage other than localised in the example) and drop variables not required (to reduce I/O time).

```
/* Restrict to localised */
if stage=1;

/* Create a unique ID for each observation */
id+1;

/* Add 0.5 to all survival times */
surv_mm = surv_mm + 0.5;

/* The lexis macro requires a variable containing the time at entry */
entry=0;

/* an indicator variable for death due to any cause */
if status in (1,2) then d=1;
else d=0;
```

At this step we also create a unique subject identifier and create an indicator variable for death due to any cause (which must be a 0/1 dummy variable because we sum this to obtain the total number of deaths). The ID variable is not mandatory (and not referenced anywhere else in the code) so existing ID variables do not need to be renamed. The death indicator, however, is required. If such an indicator with a different name already exists then it is recommended that it be renamed to `D`.

I'll demonstrate the approach based on the following three individuals. The variable `SURV_MM` represents survival time in months, `D` is the event indicator, `AGE` is the age at diagnosis, and `YYDX` the year of diagnosis.

```
  ID     SEX      AGE    YYDX    SURV_MM    D
   2    Male       80     80         8.5    1
  99    Female     77     79        31.5    1
4999    Male       80     92        46.5    0
```

4.0.1 Splitting time at risk using the lexis macro

We now use the `lexis` macro to split the observation time for each individual into multiple observations, one for each band of follow-up time. In the example we will split the observations into intervals of one year with a maximum of 10 years. That is, we wish to create a life table with annual intervals up to 10 years of follow-up. The call to the lexis macro is as follows

```
%lexis (
data=&individ.,
out=&individ.,
breaks = %str( 0 to 10 by 1 ),
origin = 0,
entry = entry,
exit = surv_mm,
fail = d,
scale = 12,
right = right,
risk = y,
lrisk = ln_y,
lint = length,
cint = w,
nint = fu
)
;
```

The lexis macro was written by Bendix Carstensen and available from his web site [2] along with example and help files. I modified the code to include variables required for life table estimates of survival (e.g., `lint`, `cint`, `nint`, and `right`). Since these variables are not output by default they need to be specified on the macro call. Both the underlying philosophy and the syntax of the lexis macro are similar to the Stata `stsplit` command. In this framework, patients enter at time zero (which is also the time origin) and exit at the time specified in `surv_mm`. For estimating relative survival the timescale must be in years (i.e., it should be transformed using the `scale` parameter if the original scale is in other units) and the breakpoints must be specified in years (since the program assumes time at risk is given in years). Since survival time in our data is recorded in months, we specify `scale = 12`.

The life table intervals do not have to be of integer length and neither do they all have to be of equal lengths. For example, either of the following could be used

```
breaks = %str( 0,0.5,1,2,3,4,5,10,20,30 )
```

or

```
breaks = %str( 0 to 10 by 0.5 ).
```

4.0.2 Specifying dates of entry and exit rather than the survival time

Rather than using the information on survival time in the variable surv_mm we can, alternatively, specify the dates of diagnosis and exit. That is, individuals enter the study at the date of diagnosis (which is the time origin) and exit at the date of death or censoring. The origin parameter is required since the time origin (i.e., time zero) occurs on a different date for each individual. Since the underlying time unit is now days we specify scale=365.25 to transform to years.

```
%lexis (
data=&individ.,
out=&individ.,
breaks = %str( 0 to 10 by 1 ),
origin = dx,
entry = dx,
exit = exit,
fail = d,
scale = 365.25,
right = right,
risk = y,
lrisk = ln_y,
lint = length,
cint = w,
nint = fu
)
;
```

The Finnish cancer registry records only the month and year of diagnosis so the SAS date variables dx and exit have been estimated from available information. In our example the resulting life tables are identical for the two alternative methods of defining the time at risk although, in practice, it is possible to obtain slightly different results.

4.0.3 Updating age and period after splitting

Splitting the data results in the following

```
  ID   SEX      AGE    YYDX    D   W  LEFT   FU       Y     LENGTH

   2   Male      80     80     1   0     0    1   0.70833     1
  99   Female    77     79     0   0     0    1   1.00000     1
  99   Female    77     79     0   0     1    2   1.00000     1
  99   Female    77     79     1   0     2    3   0.62500     1
4999   Male      80     92     0   0     0    1   1.00000     1
4999   Male      80     92     0   0     1    2   1.00000     1
4999   Male      80     92     0   0     2    3   1.00000     1
4999   Male      80     92     0   1     3    4   0.87500     1
```

The variable LEFT holds the left breakpoint of the interval and Y the time at risk during the interval. LENGTH is the length of the interval (as distinct from the time at risk during the interval which is given by Y). The units for each of these variables is years.

Note that the variables AGE and YYDX represent the age and year at diagnosis, not the attained age and year during the interval. We now want to create variables for attained age and calendar year which are 'updated' for each observation for a single individual. These are the variables by which we will merge in the expected probabilities of death, so they must have the same names and same format as the variables indexing the POPMORT file (sex, _year, and _age in this example).

```
data &individ;
set &individ;
_age=floor(age+left);
_year=floor(yydx+left);
run;
```

This results in the following:

```
  ID  SEX     AGE   _AGE  YYDX  _YEAR  D  W  FU       Y    LENGTH

   2  Male    80     80    80   1980   1  0   1  0.70833     1
  99  Female  77     77    79   1979   0  0   1  1.00000     1
  99  Female  77     78    79   1980   0  0   2  1.00000     1
  99  Female  77     79    79   1981   1  0   3  0.62500     1
4999  Male    80     80    92   1992   0  0   1  1.00000     1
4999  Male    80     81    92   1993   0  0   2  1.00000     1
4999  Male    80     82    92   1994   0  0   3  1.00000     1
4999  Male    80     83    92   1995   0  1   4  0.87500     1
```

Note that we must keep track of both age at diagnosis (which is used, for example, to stratify life tables) and attained age (upon which the expected probability of death depends). Similarly, we must keep track of both year of diagnosis and 'attained' year. My suggestion is that the variable names for the updated quantities be prefixed with an underscore.

4.0.4 Merging with the popmort file to get the expected survival proportions

The next step is to merge in the expected survival proportions. Both the data file and the file of expected survival proportions are indexed by `sex`, `_year`, and `_age`.

```
data &individ;
length d w fu 4 y ln_y length 5;
merge &individ(in=a) &popmort(in=b);
by sex _year _age;
if a;
/* Need to adjust for interval lengths other than 1 year */
p_star=prob**length;
/* Expected number of deaths */
d_star=-log(p_star)*(y/length);
run;
```

The population mortality files contains probabilities of surviving one completed year. If the life table interval is of length other than one year then the probability of surviving from the start to the end of the interval must be calculated. If the probability of surviving one year from a given date is p then the probability of surviving k years (where k is any real number) is p^k (provided k is not much larger than 1). For example, if $k = 0.5$ then we have $p^{0.5} = \sqrt{p}$. Here you can see why we need the population mortality data to be specified in the form of annual probabilities of survival and we need the intervals to be specified in units of years.

The expected number of deaths is equal to the expected cumulative hazard for the interval multiplied by the proportion of the interval for which the patient was at risk. The expected cumulative hazard is given by $-\ln(p^*)$ where p^* is the expected survival probability.

After merging in the expected probabilities of surviving (`P_STAR`) we have the following

```
  ID  SEX     AGE   _AGE  YYDX  _YEAR  D  W  FU      Y     LENGTH   P_STAR

   2  Male     80    80    80   1980   1  0   1  0.70833     1     0.88573
  99  Female  77     77    79   1979   0  0   1  1.00000     1     0.94384
  99  Female  77     78    79   1980   0  0   2  1.00000     1     0.93809
  99  Female  77     79    79   1981   1  0   3  0.62500     1     0.93755
4999  Male     80    80    92   1992   0  0   1  1.00000     1     0.90338
4999  Male     80    81    92   1993   0  0   2  1.00000     1     0.89360
4999  Male     80    82    92   1994   0  0   3  1.00000     1     0.88628
4999  Male     80    83    92   1995   0  1   4  0.87500     1     0.87186
```

4.0.5 Collapsing the data and constructing life table estimates of survival

We now collapse the data as the first step in constructing life table estimates of survival.

```
proc summary data=&individ nway;
var d w p_star y d_star;
id range length;
class &vars fu; /* Follow-up must be the last variable in this list */
output out=&grouped(drop=_type_ rename=(_freq_=l))
              sum(d w y d_star)=d w y d_star mean(p_star)=p_star;
format &formats ;
run;
```

Our data set now contains one observation for each life table interval and includes variables for number alive at the start of the interval (`l`), number of deaths during the interval (`d`), number of patients whose survival time was censored during the interval (`w`), and expected survival proportion for the interval (`p_star`). The next step is to calculate the interval-specific observed and relative survival and then estimate cumulative survival by taking the product of the interval-specific estimates.

```
data &grouped;
retain cp cp_star cr 1;
set &grouped;
if fu=1 then do;
  cp=1; cp_star=1; cr=1; se_temp=0;
  end;
l_prime=l-w/2;
ns=l_prime-d;
/* Two alternative approaches to estimating interval-specific survival */
/* Must use the hazard approach for period analysis */
p=exp(-(d/y)*length); /* transforming the hazard */
p=1-d/l_prime; /* actuarial approach */
r=p/p_star;
cp=cp*p;
cp_star=cp_star*p_star;
cr=cp/cp_star;
ln_y_group=log(l_prime-d/2);
ln_y=log(y);
d_star_group=l_prime*(1-p_star);
excess=(d-d_star)/y;
se_p=sqrt(p*(1-p)/l_prime);
se_r=se_p/p_star;
se_temp+d/(l_prime*(l_prime-d)); /* Component of the SE of the cumulative survival */
se_cp=cp*sqrt(se_temp);
se_cr=se_cp/cp_star;
... code for confidence intervals ...
run;
```

The final step is to print the estimates (using PROC PRINT) in the form of life tables. A list of variables available is provided in Section 4.6 (page 15).

```
Colon carcinoma diagnosed in Finland 1975-1994 (follow-up to 1995)
Life table estimates of patient survival
The Ederer II method is used to estimate expected survival

Sex=Male Year of diagnosis=1975-84 Age at diagnosis=0-44

                 Interval-              Interval-            Interval-
                 specific  Cumulative  specific  Cumulative specific  Cumulative
                 observed   observed   expected   expected  relative   relative
Int.     N  D  W  survival   survival   survival   survival  survival   survival

0 - 1   75  4  0  0.94667    0.94667    0.99697    0.99697   0.94954    0.94954
1 - 2   71  8  0  0.88732    0.84000    0.99682    0.99381   0.89015    0.84524
2 - 3   63  1  1  0.98400    0.82656    0.99649    0.99032   0.98747    0.83464
3 - 4   61  3  0  0.95082    0.78591    0.99625    0.98660   0.95440    0.79658
4 - 5   58  3  0  0.94828    0.74526    0.99601    0.98266   0.95208    0.75841
5 - 6   55  2  0  0.96364    0.71816    0.99562    0.97836   0.96787    0.73404
6 - 7   53  0  0  1.00000    0.71816    0.99532    0.97378   1.00470    0.73749
7 - 8   53  0  0  1.00000    0.71816    0.99491    0.96882   1.00512    0.74127
8 - 9   53  1  0  0.98113    0.70461    0.99453    0.96352   0.98653    0.73128
9 - 10  52  2  0  0.96154    0.67751    0.99418    0.95792   0.96717    0.70727
```

4.1 Checking the estimates using PROC LIFETEST

The estimates of observed survival obtained using my approach should be identical to the estimates obtained from PROC LIFETEST - it is recommended that such a test be performed. For example, the life estimates shown in the previous table can be obtained with the following code:

```
proc lifetest data=colon.colon(where=(stage=1)) method=act width=12;
time surv_mm*status(0,4);
strata sex yydx age;
format sex sex. age age. yydx yydx.;
run;
```

```
Stratum 16: SEX = Male  YYDX = 1975-84  AGE = 0-44
                                                                Conditional
                                          Effective  Conditional  Probability
    Interval        Number   Number      Sample    Probability    Standard
 [Lower,     Upper)  Failed  Censored      Size     of Failure       Error    Survival

        0         12      4         0      75.0        0.0533       0.0259      1.0000
       12         24      8         0      71.0        0.1127       0.0375      0.9467
       24         36      1         1      62.5        0.0160       0.0159      0.8400
       36         48      3         0      61.0        0.0492       0.0277      0.8266
       48         60      3         0      58.0        0.0517       0.0291      0.7859
       60         72      2         0      55.0        0.0364       0.0252      0.7453
       72         84      0         0      53.0             0            0      0.7182
       84         96      0         0      53.0             0            0      0.7182
       96        108      1         0      53.0        0.0189       0.0187      0.7182
      108        120      2         0      52.0        0.0385       0.0267      0.7046
      120        132      1         0      50.0        0.0200       0.0198      0.6775
```

Note how PROC LIFETEST presents the interval-specific failure (rather than survival) proportion and that the estimates of cumulative survival are presented for the start of each interval (i.e., they are offset by one row compared to the way in which I present the estimates).

4.2 Standard error of the observed survival proportion

The most widely used method for estimating the standard error of the estimated survival proportion is the method described by Greenwood (1926) [3] and it is this method I have used. The standard error of the cumulative (observed or cause-specific) survival proportion up until the end of interval i, denoted ${}_1p_i$, is given by

$$\mathrm{SE}({}_1p_i) = {}_1p_i \left[\sum_{j=1}^{i} \frac{d_j}{l'_j(l'_j - d_j)} \right]^{\frac{1}{2}}, \tag{1}$$

where l'_i is the effective number at risk at the start of interval i and d_i the number of deaths during interval i. Non-integer values for l'_i, e.g. $l'_i = 20.5$, do not cause any problems in practical use. The SAS code for calculating the standard error is

```
se_temp+d/(l_prime*(l_prime-d));
se_cp=cp*sqrt(se_temp);
```

The first line uses the 'Sum' statement to calculate the summation within square brackets in Equation 1.

For a single interval, Equation 1 reduces to

$$\mathrm{SE}(p_i) = p_i \left\{ \frac{d_i}{l'_i(l'_i - d_i)} \right\}^{\frac{1}{2}} = \sqrt{p_i(1 - p_i)/l'_i},$$

which is the familiar binomial formula for the standard error of the observed interval-specific survival proportion based on l'_i trials. It can also be shown that, in the absence of censoring, Equation 1 reduces to the binomial standard error.

4.3 Standard error of the relative survival ratio

The variance of the expected survival proportion is very small in comparison to the variance of the observed survival proportion and, in practice, it is assumed that the expected survival proportion is a fixed constant. The variance of the relative survival ratio (both interval-specific and cumulative) is then given by

$$\begin{aligned} \mathrm{var}(r) &= \mathrm{var}(p/p^*) \\ &= \mathrm{var}(p)/(p^*)^2. \end{aligned} \tag{2}$$

That is, the variance of the relative survival ratio is given by the variance of the observed survival rate, divided by the square of the expected survival rate. We have made use of the result that, for a random variable X and a constant a, $\mathrm{var}(aX) = a^2\mathrm{var}(X)$. The variance of the observed survival proportion ($\mathrm{var}(p) = \mathrm{SE}(p)^2$) is calculated using Greenwood's formula (Section 4.2). The standard error (SE) of the relative survival ratio is given by $\mathrm{SE}(r) = \mathrm{SE}(p)/p^*$.

4.4 Confidence intervals for observed and relative survival

A confidence interval for the survival function can be obtained by assuming that the estimate is normally distributed around the true value with estimated variance given by the square of the standard error (Section 4.2). A two-sided $100(1-\alpha)\%$ confidence interval ranges from $p - z_{\alpha/2}\mathrm{SE}(p)$ to $p + z_{\alpha/2}\mathrm{SE}(p)$, where p is the estimated survival (which can be an interval-specific or cumulative observed, cause-specific, or relative survival), $\mathrm{SE}(p)$ the associated standard error, and $z_{\alpha/2}$ the upper $\alpha/2$ percentage point of the standard normal distribution. For a 95% confidence interval, $z_{\alpha/2} = 1.96$, and for a 99% confidence interval, $z_{\alpha/2} = 2.58$.

As a rule of thumb, the normal approximation for a single interval i is usually appropriate when both l'_ip_i and $l'_i(1 - p_i)$ are greater than or equal to 5 [4]. Confidence intervals obtained in this way are symmetric about the point estimate and can sometimes contain implausible values for the survival rate, i.e., values less than zero or greater than one. The theoretical upper bound for the relative survival rate, which can be greater than one, is $1/p^*$, where p^* is the expected survival rate.

One method of obtaining confidence intervals for the observed survival rate in the range [0,1] is to transform the estimate to a value in the range $[-\infty, \infty]$, obtain a confidence interval on the transformed scale, and then back-transform the confidence interval to [0,1]. Suitable transformations are the logistic transformation, $\log[p/(1-p)]$, or the complementary log-log transformation, $\log[-\log(p)]$. I have used the complementary log-log transformation, which involves constructing the confidence intervals on the log-hazard scale. To estimate confidence intervals for the relative survival ratio using this method, we first transform the estimated cumulative observed survival proportion. We will write this transformation as $g(\text{OSR}) = \log[-\log(\text{OSR})]$, where g is the complementary log-log transformation, which transforms the cumulative observed survival rate to the log cumulative hazard scale. We also require an estimate of the variance of the observed survival proportion on the log hazard scale. Using the delta method, the variance of a function, g, of a random variable, X, can be approximated by

$$\text{var}\{g(X)\} \approx \left\{ \frac{\mathrm{d}g(X)}{\mathrm{d}X} \right\}^2 \text{var}(X) \tag{3}$$

If we denote the cumulative observed survival proportion by X then, noting that

$$\frac{\mathrm{d}\log[f(X)]}{\mathrm{d}X} = \frac{1}{f(X)} \frac{\mathrm{d}f(X)}{\mathrm{d}X}, \tag{4}$$

we have

$$\text{var}\{g(X)\} = \text{var}\{\log[-\log(X)]\} \approx \frac{1}{[X \log(X)]^2} \text{var}(X). \tag{5}$$

An estimated 95% confidence interval on the log hazard scale is therefore given by $g(\text{OSR}) \pm 1.96\sqrt{\text{var}\{g(X)\}}$, which is then back-transformed to give a 95% confidence interval for the OSR. To obtain a CI for the relative survival ratio, the upper and lower confidence limits for the observed survival proportion are simply divided by the expected survival proportion.

4.5 Tabular and graphical presentation of survival estimates

Code illustrating how estimates can be presented in tabular of graphical format is available at:
`http://www.pauldickman.com/rsmodel/table_of_survival_estimates.sas`
`http://www.pauldickman.com/rsmodel/graph_of_survival_estimates.sas`

4.6 Variables contained in the output data sets GROUPED and INDIVID

Contents of the data set GROUPED

RANGE	Text label for interval
FU	Index for the interval (1,2,3,...)
LENGTH	Interval length (k_i)
L	Number alive at the start of the interval (l_i)
L_PRIME	Effective number at risk ($l'_i = l_i - w_i/2$)
D	Deaths during the interval (d_i)
W	Withdrawals (censorings) during the interval (w_i)
P	Interval-specific observed survival (p_i)
P_STAR	Interval-specific expected survival (p^*_i)
R	Interval-specific relative survival (r_i)
CP	Cumulative observed survival
CP_STAR	Cumulative expected survival
CR	Cumulative relative survival
Y	Person-time (years) at risk (using ‘exact’ times) (y_i)
NS	Number surviving the interval ($l'_i - d_i$) (outcome variable in the binomial model)
LN_Y_GROUP	ln(person-time) calculated as $\ln(l'_i - d_i/2)$
LN_Y	ln(person-time) using ‘exact’ person-time (i.e., $\ln(y_i)$)
D_STAR	Expected deaths using ‘exact’ times ($d^*_i = -\ln(p^*_i)y_i/k_i$)
D_STAR_GROUP	Expected deaths for grouped data ($d^*_i = l'_i(1 - p^*_i)$)
EXCESS	Empirical excess hazard ($= (d_i - d^*_i)/y_i$)
SE_P	Standard error of P
SE_R	Standard error of R
SE_CP	Standard error of CP
SE_CR	Standard error of CR
LO_P	Lower 95% CI for P
HI_P	Upper 95% CI for P
LO_R	Lower 95% CI for R
HI_R	Upper 95% CI for R
LO_CP	Lower 95% CI for CP
HI_CP	Upper 95% CI for CP
LO_CR	Lower 95% CI for CR
HI_CR	Upper 95% CI for CR
SEX	Sex
YYDX	Year of diagnosis (in categories defined by a format)
AGE	Age at diagnosis (in categories defined by a format)

Contents of the data set INDIVID

RANGE	Text label for interval
FU	Index for the interval (1,2,3,...)
LENGTH	Interval length (k_i)
D	Indicator for death during the interval (d_i)
W	Indicator for withdrawal (censoring) during the interval (w_i)
P_STAR	Interval-specific expected survival (p^*_i)
Y	Person-time (years) at risk (using ‘exact’ times) (y_i)
LN_Y	ln(person-time) using ‘exact’ person-time (i.e., $\ln(y_i)$)
D_STAR	Expected deaths using ‘exact’ times ($d^*_i = -\ln(p^*_i)y_i/k_i$)
SEX	Sex
YYDX	Year of diagnosis (in categories defined by a format)
AGE	Age at diagnosis (in categories defined by a format)

5 Modelling relative survival using SAS

The code in `survival.sas` produces two output data sets, one containing individual data and one containing grouped data. The file `models.sas` contains SAS code for estimating a model for relative survival using five different approaches:

1. Grouped data, GLM with binomial error structure (Hakulinen-Tenkanen approach)
2. Grouped data, GLM with Poisson error structure (i.e., Poisson regression)
3. Exact survival times, Poisson regression estimated using collapsed data
4. Exact survival times, Poisson regression estimated using individual level data
5. Exact survival times, full likelihood (Estève *et al.* approach)

5.1 Description of the model

When modelling relative survival, the hazard function at time since diagnosis t for persons diagnosed with cancer (with covariate vector $\mathbf{z}$) is modelled as the sum of the expected hazard, $\lambda^*(t;\mathbf{z})$, and the excess hazard due to a diagnosis of cancer, $\nu(t;\mathbf{z})$. That is,

$$\lambda(t;\mathbf{z}) = \lambda^*(t;\mathbf{z}) + \nu(t;\mathbf{z}). \tag{6}$$

The expected hazard is annotated with an asterisk to indicate that it is estimated from external data (general-population mortality rates) as opposed to, for example, the baseline hazard in a Cox proportional hazards model [5], an arbitrary function which is not estimated. Some authors prefer to write the expected hazard as $\lambda^*(t;\mathbf{z_1})$, where $\mathbf{z_1}$ is a subvector of $\mathbf{z}$, in order to indicate that the expected hazard is generally assumed to depend only on a subset of the covariates available (typically age, sex, and period). The expected hazard does not depend, for example, on tumour-specific covariates such as histology or stage. We will write, for simplicity, that the expected hazard is a function of $\mathbf{z}$, even though it does not vary over all elements of $\mathbf{z}$.

The model is known as an additive hazards model or a relative survival model, since it can be written as

$$S(t;\mathbf{z}) = S^*(t;\mathbf{z}) \times r(t;\mathbf{z}) \tag{7}$$

where $S(t;\mathbf{z})$, $S^*(t;\mathbf{z})$, and $r(t;\mathbf{z})$ represent cumulative observed, expected, and relative survival. For population-based cancer survival data, such models are generally biologically more plausible and provide a better fit to the data than multiplicative models [6, 7, 8, 9]. The hazards are assumed to be constant within pre-specified subintervals (bands) of follow-up time (i.e. piecewise constant hazards). These intervals are typically of length one year, although it is common to use shorter intervals early in the follow up (e.g. during the first year) and longer intervals later in the follow-up (e.g. after 10 years). A set of indicator variables are constructed (one indicator variable for each interval excluding the reference interval) and incorporated into the covariate vector. We will use $\mathbf{x}$ to denote the covariate vector that contains indicator variables for these bands of follow-up time in addition to the other covariates $\mathbf{z}$. Our primary interest is in the excess hazard component, ν, which is assumed to be a multiplicative function of the covariates, written as $\exp(\mathbf{x}\beta)$. The basic relative survival model is therefore written as

$$\lambda(\mathbf{x}) = \lambda^*(\mathbf{x}) + \exp(\mathbf{x}\beta). \tag{8}$$

Parameters representing the effect in each follow-up interval are estimated in the same way as parameters representing the effect of, for example, age, sex, or histology. Implicit in Equation 8 is the assumption that the excess hazards for any two patient subgroups are proportional over follow-up time. Non-proportional excess hazards can, however, be incorporated by including time by covariate interaction terms in the model. The exponentiated parameter estimates have an interpretation as excess hazard ratios, sometimes known as relative excess risks [10]. An excess hazard ratio of, for example, 1.5 for males compared to females implies that the excess hazard associated with a diagnosis of cancer is 50% higher for males than females.

Various approaches to estimating the model, and their implementation in SAS, are described in the following sections.

5.2 Estève *et al.* full likelihood approach

Estève *et al.* [8] described a method for estimating the model in Equation 8 directly from individual-level data using a full maximum likelihood approach. The likelihood function is

$$L = \prod_{i=1}^{n} \exp(-\int_0^{t_i} \lambda(s)\, ds)[\lambda(t_i)]^{d_i}, \tag{9}$$

where t_i is the survival time and d_i the failure indicator variable (1 if t_i is the time of death; 0 if the survival time is censored at t_i) for each of the $i = 1, \ldots, n$ individuals.

Writing the total hazard as the sum of the expected hazard and the excess hazard, the log-likelihood function is

$$l(\beta) = -\sum_{i=1}^{n} \int_0^{t_i} \lambda^*(s)\, ds - \sum_{i=1}^{n} \int_0^{t_i} \nu(s)\, ds + \sum_{i=1}^{n} d_i \ln[\lambda^*(t_i) + \nu(t_i)]. \tag{10}$$

Although the model is specified in continuous time it is assumed, as with all approaches described here, that the hazard is constant within pre-specified bands of time and the excess hazard $\nu(t)$ is written as $\exp(\mathbf{x}\beta)$. Estimation of the model is simplified if each observation is split into separate observations for each band of follow-up. Rather than evaluating the log likelihood for each subject and summing over subjects (the Estève *et al.* approach) we evaluate the log-likelihood for each subject-band. The log likelihood function, expressed in terms of the J subject-band observations, is

$$l(\beta) = \sum_{j=1}^{J} \left[d_j \ln[\lambda^*(\mathbf{x}_j) + \exp(\mathbf{x}_j\beta)] - y_j \exp(\mathbf{x}_j\beta)\right]. \tag{11}$$

This model can be estimated based on the 'individual-level data' output from `survival.sas` using PROC NLP (part of SAS/OR) — we simply specify the log likelihood function (Equation 11) in terms of the parameters and have SAS maximise it for us.

```
proc nlp data=nlp_data cov=2 vardef=n;
title2 'Full likelihood estimation from individual data (Esteve approach)';
max loglike;
parms int fu_2-fu_5 female year2 age2-age4;
theta = int+fu_2*fu2+fu_3*fu3+fu_4*fu4+fu_5*fu5+year2*year8594
        +age2*age_gr2+age3*age_gr3+age4*age_gr4+female*sex2;
loglike = d*log(-log(p_star)+exp(theta))-exp(theta)*y;
run;
```

Note that, in order to model categorical explanatory variables, indicator variables must be constructed in a data step (see `models.sas`). The variables `fu2`–`fu5` are indicator variables contained in the input data set whereas `fu_2`–`fu_5` are parameters to be estimated by PROC NLP. Estimates of relative excess risks are obtained by writing the parameter estimates to a data set and exponentiating the parameter estimates is a data step (see `models.sas`). Maximum likelihood estimation can also be performed using PROC NLIN (part of SAS/STAT) although my experience has been that convergence is more easily obtained using PROC NLP.

5.3 Poisson regression approach

The relative survival model (Equation 8) assumes piecewise constant hazards which implies a Poisson process for the number of deaths in each interval; see Andersen et al. [11, pp. 409] or Breslow and Day [12, Section 4.2]. The log likelihood in Equation 11 is therefore identical to the log-likelihood for grouped Poisson data with intensity $\lambda^*(\mathbf{x}) + \exp(\mathbf{x}\beta)$ [12, pp. 185] except we omitted the term $-y_j\lambda^*(\mathbf{x_j})$ since it did not depend on β.

This implies that the relative survival model can be estimated in the framework of generalised linear models using a Poisson assumption for the observed number of deaths. If the model is estimated from subject-band observations the estimates will be identical to those obtained using the full likelihood approach (Section 5.2) since we maximise the same likelihood based on the same data. We can also, however, estimate the model based on collapsed or grouped data, in which case the estimates differ slightly.

We assume that the number of deaths, d_j, for observation j can be described by a Poisson distribution, $d_j \sim \text{Poisson}(\mu_j)$ where $\mu_j = \lambda_j y_j$ and y_j is person-time at risk for the observation. The observations can represent either life table intervals (in which case there can be multiple deaths per observation), individual patients, or subject-bands (as in Section 5.2).

Equation 8 is then written as

$$\mu_j/y_j = d_j^*/y_j + \exp(\mathbf{x}\beta), \tag{12}$$

which can be written as

$$\ln(\mu_j - d_j^*) = \ln(y_j) + \mathbf{x}\beta, \tag{13}$$

where d_j^* is the expected number of deaths (due to causes other than the cancer of interest and estimated from general population mortality rates). This implies a generalised linear model with outcome d_j, Poisson error structure, link $\ln(\mu_j - d_j^*)$, and offset $\ln(y_j)$. Such a model was suggested by Berry (1983) [13] and is specified in PROC GENMOD as follows

```
proc genmod data=&grouped(where=(fu le 5)) order=formatted;
title2 'Poisson error model fitted to collapsed data';
title3 'Main effects model (follow-up, sex, age, and dgnyear)';
fwdlink link = log(_MEAN_-d_star);
invlink ilink= exp(_XBETA_)+d_star;
class fu sex age yydx;
model d = fu sex yydx age  / error=poisson offset=ln_y type3;
format fu fu. age age. yydx yydx.;
run;
```

Estimates of relative excess risks are obtained by writing the parameter estimates to a data set and exponentiating the parameter estimates is a data step (see `models.sas`).

5.3.1 Estimation based on collapsed or grouped data

The model can be estimated directly from subject-band observations (in the data set `&individ`) or the subject-band observations can be collapsed to give one observation for each covariate pattern (d, d^*, and y are summed within each covariate pattern) as in the data set `&grouped`. Estimating a standard Poisson regression model (with logarithmic link and offset $\ln(y_j)$) gives identical estimates for both individual and collapsed data. Estimating Equation 13 based on collapsed data, however, leads to slightly different estimates to those obtained from subject-band observations since d^* varies within each covariate pattern (i.e. combination of follow-up interval, sex, period, age group, etc.). The model can also be estimated based on grouped data (information is available only on the number of events in a time interval rather than exact times to event for each individual). In this case we can estimate the expected number of deaths as $d_i^* = l_i'(1 - p_i^*)$ and the log person-time at risk as $\ln(l_i' - d_i/2)$. These two quantities are stored in the variable `D_STAR_GROUP`) and `LN_Y_GROUP` in the GROUPED data set.

5.4 Hakulinen–Tenkanen approach

Hakulinen and Tenkanen [7] estimate the relative survival model from grouped survival data using an assumption that the number of deaths in each life table interval can be modelled using a binomial distribution. The model is estimated in the framework of generalised linear models [14] where the outcome is $l'_{ki} - d_{ki}$ (the number of patients surviving the interval), the error structure binomial with denominator l'_{ki}, and the link function complementary log-log combined with a division by p^*_{ki}. That is,

$$\ln\left[-\ln\frac{p_{ki}}{p^*_{ki}}\right] = \mathbf{x}\beta. \tag{14}$$

Such a model is specified in PROC GENMOD as follows

```
proc genmod data=&grouped(where=(fu le 5)) order=formatted;
title2 'Binomial error model fitted to grouped data';
title3 'Main effects model (follow-up, sex, age, and dgnyear)';
fwdlink link = log(-log(_mean_/p_star));
invlink ilink = exp(-exp(_xbeta_))*p_star;
class fu sex age yydx;
model ns/l_prime = fu sex yydx age / error=bin type3;
format fu fu. age age. yydx yydx.;
run;
```

Estimates of relative excess risks are obtained by writing the parameter estimates to a data set and exponentiating the parameter estimates in a data step (see `models.sas`).

This approach is not preferred for modelling period estimates, especially in my implementation since w and l' in my code do not take account of the number of individuals whose survival time is left truncated.

6 Estimating and modelling relative survival in SAS using period analysis

In 1996 Hermann Brenner [15, 16] suggested that lifetable estimates of patient survival could be made using a period rather than the traditional cohort approach. Period analysis has since been shown to provide more accurate predictions of newly diagnosed patients and is able to detect temporal trends in patient survival sooner than the traditional cohort approach [17, 18, 19].

Professor Brenner and colleagues have produced SAS macros for estimating survival using period analysis which are available at
`http://www.imbe.med.uni-erlangen.de/issan/SAS/period/period.htm`

These macros can be used to estimate survival using either the cohort or the period approach and expected survival can be calculated using either the Ederer II or the Hakulinen method. My approach to estimating survival can also be used for period analysis and excess mortality can be modelled in the same manner as with cohort analysis. I estimate expected survival using the Ederer II method (the Hakulinen method is not implemented) although this is sufficient when the primary interest is in modelling.

6.1 Overview of my approach to period analysis in SAS

There are two differences when using my approach for period analysis compared to cohort analysis. The first is that we use the `lexis` macro an additional time to restrict the time at risk to the calendar window of interest. The second difference is that interval-specific observed survival is estimated by transforming the estimated hazard rate rather than by using the actuarial estimator. The approach is as follows:

1. Split the data by calendar time using the `lexis` macro so, for each individual, we retain only the time at risk during the calendar window of interest. For example, we might be interested in the time at risk between 1 January 1994 and 31 December 1995. If an individual is not at risk during the period of interest (e.g., they have died before the start of the window or are diagnosed after the window) then they do not contribute to the analysis. After this step our data set will consist of (at most) one observation per individual.
2. Split the resulting data by time since diagnosis in the usual manner. Each individual may contribute more than one observation.
3. Collapse the data and estimate life tables in the usual manner. A slight difference from the cohort approach is that we estimate survival by transforming the cumulative hazard (rather than using the actuarial method).

6.2 Estimating survival by transforming the hazard

The standard actuarial estimator for the interval-specific observed survival for interval i is

$$p_i = (1 - d_i/l'_i)$$

where d_i is the number if deaths in the interval and $l'_i = 1 - w_i/2$ is the 'effective number at risk' (w_i is the number censored during the interval). In period analysis survival times can be left truncated in addition to being right censored so fewer subjects are at risk for the full interval. As such, w_i would need to represent the number of individuals whose survival time was left truncated or right censored. However, it is possible that some survival times would be both left truncated and right censored during the interval (and hence be, on average, at risk for only one quarter of the interval) so the estimator would need to be modified to account for this.

An alternative approach is to use the relationship that the survivor function is equal to the exponential of the negative of the cumulative hazard ($S = \exp(-\Lambda)$). The cumulative hazard, Λ, is the area under the hazard curve. We can estimate the average hazard for the interval as $\lambda_i = d_i/y_i$ where d_i is the number of deaths and y_i the person-time at risk in the interval. If the hazard is assumed to be constant at this value during the interval then the cumulative hazard for the interval is $\Lambda_i = k_i \times d_i/y_i$ where k_i is the width of the interval. Our estimate of the interval-specific observed survival is therefore

$$p_i = \exp(k_i \times -d_i/y_i)$$

Since this approach assumes the hazard is constant within the interval, it is sensitive to the choice of interval length, unlike the actuarial approach which gives the same estimates of cumulative observed survival independent of the choice of intervals.

6.3 An example

Consider the following 7 individuals diagnosed with colon carcinoma.

```
 ID        DX         ENTRY       EXIT       D

1203     07FEB80     07FEB80     22MAY83     1
5128     07JUN92     07JUN92     22MAR93     1
5150     07JUN92     07JUN92     22DEC95     0
5159     07JUN92     07JUN92     22AUG95     1
5647     07OCT93     07OCT93     22DEC95     0
6259     07APR94     07APR94     22DEC94     1
6260     07MAY94     07MAY94     22DEC95     0
```

We wish to estimate relative survival using the period approach with a window between 1 January 1994 and 31 December 1995. We need to keep track of both the date of diagnosis and the date of entry (the date at which the individual became at risk). By default, the date of entry is the date of diagnosis but in period analysis patients are not necessarily at risk from the date of diagnosis. In our example, individuals are only at risk between 1 January 1994 and 31 December 1995. To make this restriction. We ‘split’ the data by calendar time using the lexis macro

```
%lexis (
data=&individ.,
out=&individ.,
breaks = %str( ’01jan1994’d,’31dec1995’d ),
origin = 0,
entry = entry,
exit = exit,
fail = d
)
;
```

which results in the following data set

```
 ID        DX         ENTRY       EXIT       D

5150     07JUN92     01JAN94     22DEC95     0
5159     07JUN92     01JAN94     22AUG95     1
5647     07OCT93     01JAN94     22DEC95     0
6259     07APR94     07APR94     22DEC94     1
6260     07MAY94     07MAY94     22DEC95     0
```

Note that patients 1203 and 5128 were not at risk during the window so do not appear in the data. If we choose a window sufficiently wide that no person-time was excluded then we would obtain the usual cohort estimates. We need to keep track of both the date of diagnosis (dx) and the date at which the person entered the risk set (entry). Note, for example, that patient 5150 was diagnosed on 7 June 1992 but did not enter the risk set until 1 January 1994. At the time this patient became 'at risk' he or she was in the 2nd year subsequent to diagnosis and will not, therefore, contribute to the survival estimate for the first life table interval.

We now split by time since diagnosis (in the same manner as when we estimate survival using the cohort approach).

```
%lexis (
data=&individ.,
out=&individ.,
breaks = %str( 0 to 10 by 1 ),
origin = dx,
entry = entry,
exit = exit,
fail = d,
scale = 365.25,
right = right,
risk = y,
lrisk = ln_y,
lint = length,
cint = w,
nint = fu
)
;
```

The time origin is the date of diagnosis and each patient enters the risk set at the entry date (which is not necessarily the same as the date of diagnosis as it was in cohort analysis). The resulting data set is as follows:

```
 ID     LEFT    FU      DX        ENTRY      EXIT         Y       D

5150      1      2    07JUN92    01JAN94    07JUN94    0.43121    0
5150      2      3    07JUN92    07JUN94    07JUN95    1.00000    0
5150      3      4    07JUN92    07JUN95    22DEC95    0.54004    0
5159      1      2    07JUN92    01JAN94    07JUN94    0.43121    0
5159      2      3    07JUN92    07JUN94    07JUN95    1.00000    0
5159      3      4    07JUN92    07JUN95    22AUG95    0.20602    1
5647      0      1    07OCT93    01JAN94    07OCT94    0.76454    0
5647      1      2    07OCT93    07OCT94    07OCT95    1.00000    0
5647      2      3    07OCT93    07OCT95    22DEC95    0.20671    0
6259      0      1    07APR94    07APR94    22DEC94    0.70910    1
6260      0      1    07MAY94    07MAY94    07MAY95    1.00000    0
6260      1      2    07MAY94    07MAY95    22DEC95    0.62628    0
```

The variable risk holds the time at risk during the interval. These data are then collapsed to obtain the life table.

```
                     Interval-             Interval-             Interval-
                     specific  Cumulative specific  Cumulative specific  Cumulative
                     observed   observed  expected   expected  relative   relative
Interval   N  D  W   survival   survival  survival   survival  survival   survival

0.0 - 1.0  3  1  0    0.66747    0.66747   0.94097    0.94097   0.70934    0.70934
1.0 - 2.0  4  0  1    1.00000    0.66747   0.93705    0.88173   1.06718    0.75700
2.0 - 3.0  3  0  1    1.00000    0.66747   0.91256    0.80464   1.09581    0.82953
3.0 - 4.0  2  1  1    0.26175    0.17471   0.95752    0.77046   0.27336    0.22676
```

7 The Finnish Cancer Registry

The Finnish Cancer Registry is population-based and covers the whole of Finland (population 5.1 million). The Registry was established in 1952, with 1953 being the first calendar year with complete registration. The Registry obtains information from many different sources: hospitals and other institutions with inpatient beds, physicians working outside hospitals, dentists, and pathological and cytological laboratories. The Finnish Cancer Registry also receives copies of all death certificates where cancer is mentioned. Notification of new cancer cases to the Cancer Registry is mandatory by law. If the reported information is deficient or contradictory, requests are sent to informants in order to ensure accuracy in the following areas: patient details, the primary site of the tumour, and the date of diagnosis.

The diseases registered at the Finnish Cancer Registry include, in addition to all clearly malignant neoplasms, carcinoma in situ lesions (except those of the skin), all neoplasms of the intracranial space and spinal canal irrespective of their malignancy, benign papillomas of the urinary organs, semimalignant tumours of the ovary, basal cell carcinomas of the skin, and cases of polycythaemia vera and myelofibrosis.

Various check-ups have shown that the coverage of the Cancer Registry file is almost complete with respect to cancer cases diagnosed in the Finnish population [20, 21]. All independent primary neoplasms in the same person are registered separately. When evaluating whether a new tumour is an independent cancer or a recurrence, attention is focused on, among other aspects, the time interval between the tumours, histology, and knowledge of the general behaviour of each cancer type. In principle, multiple metachronous tumours in the same organ (e.g., in the colon or skin) are registered separately, especially when they have different histologies. However, each case is evaluated individually and a primary site code ‘multiple cancer’ is also available for some organs. The International Classification of Diseases Volume 7 (ICD-7) is used at the Finnish Cancer Registry. Further details of the registry can be found in the annual incidence publications [22].

References

[1] Dickman PW, Sloggett A, Hills M, Hakulinen T. Regression models for relative survival. *Statistics in Medicine* 2004;**23**:51–64.

[2] Carstensen B. Lexis macro for splitting person-time in sas, 2004. `http://www.biostat.ku.dk/~bxc/Lexis/`.

[3] Greenwood M. *The Errors of Sampling of the Survivorship Table*, vol. 33 of *Reports on Public Health and Medical Subjects*. London: Her Majesty's Stationery Office, 1926.

[4] Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991.

[5] Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* 1972;**34**:187–220.

[6] Buckley JD. Additive and multiplicative models for relative survival rates. *Biometrics* 1984; **40**:51–62.

[7] Hakulinen T, Tenkanen L. Regression analysis of relative survival rates. *Applied Statistics* 1987;**36**:309–317.

[8] Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine* 1990;**9**:529–538.

[9] Bolard P, Quantin C, Estève J, Faivre J, Abrahamowicz M. Modelling time-dependent hazard ratios in relative survival: Application to colon cancer. *Journal of Clinical Epidemiology* 2001; **54**:986–996.

[10] Suissa S. Relative excess risk: An alternative measure of comparitive risk. *American Journal of Epidemiology* 1999;**150**:279–282.

[11] Andersen PK, Borgan , Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer-Verlag, 1995.

[12] Breslow NE, Day NE. *Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies*. IARC Scientific Publications No. 82. Lyon: International Agency for Research on Cancer, 1987.

[13] Berry G. The analysis of mortality by the subject-years method. *Biometrics* 1983;**39**:173–184.

[14] McCullagh P, Nelder JA. *Generalized Linear Models*. London: Chapman and Hall, 2nd edn., 1989.

[15] Brenner H, Gefeller O. An alternative approach to monitoring cancer patient survival. *Cancer* 1996;**78**:2004–2010.

[16] Brenner H, Gefeller O. Deriving more up-to-date estimates of long-term patient survival. *Journal of Clinical Epidemiology* 1997;**50**:211–216.

[17] Brenner H, Gefeller O, Stegmaier C, Ziegler H. More up-to-date monitoring of long-term survival rates by cancer registries: an empirical example. *Methods Inf Med* 2001;**40**:248–52.

[18] Brenner H, Hakulinen T. Up-to-date long-term survival curves of patients with cancer by period analysis. *Journal of Clinical Oncology* 2002;**20**:826–832.

[19] Brenner H, Gefeller O, Hakulinen T. Period analysis for 'up-to-date' cancer survival data: theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer* 2004;**40**:326–35.

[20] Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. Experience in Finland. *Acta Oncologica* 1994;**33**:365–369.

[21] Hakulinen T. Health care system, cancer registration and follow-up of cancer patients in Finland. In: Berrino F, Sant M, Verdecchia A, Capocaccia R, Hakulinen T, Estève J, eds., *Survival of Cancer Patients in Europe: The EUROCARE Study*, IARC Scientific Publications No. 132. Lyon: International Agency for Research on Cancer, 1995; 53–54.

[22] Finnish Cancer Registry. *Cancer Incidence in Finland 1995*. Cancer Society of Finland Publication No. 58. Helsinki: Cancer Society of Finland, 1997.