The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–24

Estimating and modelling relative survival

Paul W. Dickman
Karolinska Institutet
Stockholm, Sweden

Enzo Coviello
Department of Prevention ASL BAT/1
Minervino Murge, Italy

Michael Hills
Retired

Abstract. Relative survival, the survival analogue of excess mortality, is the method of choice for estimating patient survival using data collected by population-based cancer registries. The relative survival ratio is typically estimated from life tables as the ratio of the observed survival of the patients (where all deaths are considered events) to the expected survival of a comparable group from the general population. This article describes the command `strs` for life table estimatation of relative survival. Three methods of estimating expected survival are available and estimates can be made using either a cohort or period approach. Excess mortality can be modelled using a range of approaches including full likelihood (using the `ml` command) and Poisson regression (using the `glm` command with a user-specified link function).

Keywords: st0001, excess mortality, relative survival, survival analysis, Poisson regression, life table, cancer survival, period analysis

1 Introduction

Relative survival is the method of choice for estimating patient survival using data collected by population-based cancer registries although its utility is not restricted to studying cancer (Dickman and Adami 2006; Dickman et al. 2004). Estimating cause-specific mortality (and its analogue cause-specific survival) using cancer registry data is problematic because information on cause-of-death is often unreliable or unavailable (Gamel and Vogel 2001). We instead estimate the net mortality associated with a diagnosis of cancer in terms of excess mortality, the difference between the total mortality experienced by the patients and the expected mortality of a comparable group from the general population, matched to the patients with respect to the main factors affecting patient survival and assumed to be practically free of the cancer of interest.

Relative survival is estimated from life tables as the ratio of the observed survival of the patients (where all deaths are considered events) to the expected survival. It is usual to estimate expected survival from nationwide population life tables stratified by age, sex, calendar time, and, where applicable, race. The major advantages of relative survival are that information on cause of death is not required and that it provides a measure of the excess mortality experienced by patients diagnosed with cancer, irrespective of whether the excess mortality is directly or indirectly attributable to the cancer.

2 Methods

2.1 Estimating observed survival

For traditional cohort life tables, `strs` employs the usual actuarial estimator; 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 $l'_i = l_i - w_i/2$ is the 'effective number at risk' (w_i is the number censored during the interval). In period analysis (see Section 3.6) 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.

Whenever late entry is detected (i.e., a period approach is employed) `strs` estimates survival by transforming the estimated cumulative hazard ($S = \exp(-\Lambda)$). We can estimate the average hazard for an 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 can be 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.

2.2 Estimating expected survival

The two most widely used methods for estimating expected survival, for the purpose of estimating relative survival, are commonly known as the Ederer II method (Ederer and Heise 1959) and the Hakulinen method (Hakulinen 1982). `strs` implements both methods, Ederer II being the default, in addition to a third method that is commonly referred to as the Ederer I method (Ederer et al. 1961). Expected survival can be thought of as being calculated for a cohort of patients from the general population matched by age, sex, and period. The three methods differ regarding how long each individual is considered to be 'at risk' for the purpose of estimating expected survival.

Ederer I the matched individuals are considered to be at risk indefinitely (even beyond the closing date of the study). The time at which a cancer patient dies or is censored has no effect on the expected survival.

Ederer II the matched individuals are considered to be at risk until the corresponding cancer patient dies or is censored.

Hakulinen if the survival time of a cancer patient is censored then so is the survival time of the matched individual. However, if a cancer patient dies the matched individual is assumed to be 'at risk' until the closing date of the study.

Mathematical details of the methods are given in the appendix.

Although the Ederer I method provides unbiased estimates of the expected survival proportion, its application, together with a potentially biased observed survival proportion, results in biased estimates (usually overestimates) of the relative survival ratio (Hakulinen 1982) because the method does not allow for the fact that the potential follow-up times of the patients are of unequal length. Although the Ederer II method controls for heterogeneous observed follow-up times, the expected survival proportion is dependent on the observed mortality, leading to biased estimates (usually underestimates) of the relative survival ratio (Hakulinen 1982). Expected survival proportions estimated using the Hakulinen method are adjusted for potentially heterogeneous follow-up times among the patients and are independent of the observed mortality of the patients. A potential drawback of the Hakulinen method is that information on potential follow-up times are required for all patients. The Hakulinen method is considered slightly preferable for estimating long-term (greater than 10 years) cumulative expected survival. For estimation of interval-specific survival, which includes estimation for later modelling, there is essentially no difference between the methods.

2.3 Standard errors and confidence intervals

The standard error of the observed survival proportion is estimated using Greenwood's method (Greenwood 1926). The standard error of the relative survival ratio is estimated as the standard error of the observed survival proportion divided by the expected survival proportion (Ederer et al. 1961). This is standard practice, although Brenner and Hakulinen (2005) showed that assuming expected survival to be known (rather than estimated with random error) results in biased estimates of the standard error of the relative survival ratio (usually overestimation due to positive correlation between the standard errors of the observed and expected survival).

Confidence intervals are calculated on the log cumulative hazard scale. That is, we first calculate a confidence interval for $\log(-\log S)$ and then backtransform to the survival scale.

3 The strs command

In general, two data files are required in order to estimate relative survival; a file containing individual-level data on the patients and a file containing expected probabilities of death for a comparable general population (the 'popmort' file; see Section 3.3). The `strs` command is for use with survival-time (`st`) data; the patient data file must be `stset` using the `id()` option with time since entry in years as the timescale before using `strs`; see [ST] **stset**. The basis of the estimation algorithm is to split the data using `stsplit` thereby obtaining one observation for each individual for each life table interval (which do not have to be of equal length). The expected probabilities are then obtained by merging with the popmort file and the data collapsed to obtain one observation for each life table interval. Expected survival may be estimated using either the Ederer I (`ederer1` option), Ederer II (the default), or Hakulinen methods (`potfu` option).

3.1 Syntax

`strs using` *filename* [`if` *exp*] [`in` *range*] [`iweight=`*varname*], `breaks(`*numlist ascending*`) mergeby(`*varlist*`)` [`by(`*varlist*`) diagage(`*varname*`) diagyear(`*varname*`) attage(`*newvarname*`) attyear(`*newvarname*`) survprob(`*varname*`) maxage(`*int 99*`) standstrata(`*varname*`) brenner list(`*varlist*`) potfu(`*varname*`) format(`*%fmt*`) ederer1 notables level(`*int*`) save`[(replace)] savind(`*filename*`[, replace]) savgroup(`*filename*`[, replace])]`

`using` *filename* specifies a file containing general population survival probabilities (see Section 3.3).

Importance weights (`iweights`) can be used to produce age-standardised estimates; see the example in section 3.7.

3.2 Options

`breaks(`*numlist ascending*`)` specifies the cutpoints for the lifetable intervals as an ascending `numlist` commencing at zero. The cutpoints need not be integer nor equidistant but the units must be years, e.g., specify `breaks(0(0.0833)5)` for monthly intervals up to 5 years.

`mergeby(`*varlist*`)` specifies the variables by which the file of general population survival probabilities (the `using` file) is sorted.

`by(`*varlist*`)` specifies the life table stratification variables. One life table is estimated for each combination of these variables.

`diagage(`*varname*`)` specifies the variable containing age at diagnosis in years. Does not

have to contain integer values. Default is `age`.

`diagyear(`*varname*`)` specifies the variable containing calendar year of diagnosis. Default is `yydx`.

`attage(`*newvar*`)` specifies the variable containing attained age (i.e., age at the time of follow-up). This variable cannot exist in the patient data file (it is created as the integer part of age at diagnosis plus follow-up time) but must exist in the using file. Default is `_age`.

`attyear(`*newvar*`)` specifies the variable containing attained calendar year (i.e. calendar year at the time of follow-up). This variable cannot exist in the patient data file (it is created as the integer part of year of diagnosis plus follow-up time) but must exist in the using file. Default is `_year`.

`survprob(`*varname*`)` specifies the variable in the using file that contains the general population survival probabilities. Default is `prob`.

`maxage(`*integer*`)` 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.

`standstrata(`*varname*`)` specifies a variable defining strata across which to average the cumulative survival estimate. Weights must also be specified using `[iweight=varname]`.

`brenner` specifies that the (age) adjustment be performed using the approach proposed by Brenner et al. (2004a). This option requires that `iweight` and `standstrata()` are also specified.

`list(`*varlist*`)` specifies the variables to be listed in the life tables.

`potfu(`*varname*`)` specifies a variable containing the last time of potential follow-up. This is required for calculating Hakulinen estimates of expected survival and causes `strs` to report Hakulinen estimates by default. This variable must be in the same time units as the exit time and a variable containing the time origin must be specified; in practice, it is recommended that `potfu()` specify a variable containing a date and that the data be `stset` by specifying the dates of entry and exit with the entry date as the time origin. See the example in Section 3.5.

`format(`*%fmt*`)` specifies the format for variables containing survival estimates. Default is `%6.4f`.

`ederer1` specifies that Ederer I estimates be calculated and causes `strs` to report these by default (unless `potfu()` is also specified).

`notables` suppresses display of the life tables.

`level(`*integer*`)` sets the confidence level; default is based on the value of global macro `S_level` which, by default, takes a value of 95.

`save[(replace)]` creates two output data sets, `individ.dta` contains one observation for each patient for each life table interval and `grouped.dta` contains one observa-

tion for each life table interval. Use `save(replace)` to overwrite these files. Excess mortality (relative survival) may be modelled using these output data sets (see section 4).

`savind(filename[,replace]) savgroup(filename[,replace])` may be used to specify alternative filenames for the individual and grouped output data sets.

3.3 The population mortality file

The population mortality file (typically named `popmort.dta`) must contain general population survival probabilities (conditional probabilities of surviving one year) stratified by all variables upon which expected survival depends – typically age, sex, and period – but can also include, for example, race, region/country of residence, or social class (Coleman et al. 1999). The filename is specified via the `using` option and the `mergeby(varlist)` option specifies the variables by which the file is sorted. Following is a listing of the first five rows of the Finnish popmort file.

```
. use popmort, clear
. list in 1/5
```

	sex	_year	_age	prob
1.	1	1951	0	.96429
2.	1	1951	1	.99639
3.	1	1951	2	.99783
4.	1	1951	3	.99842
5.	1	1951	4	.99882

Probabilities must be provided for every year that the patients will attain during follow-up; if data are not available for recent years it is standard practice to assume the probabilities are the same as those most recently available (`strs` does not do this automatically, the popmort file must be extended). Patient survival is often estimated for subgroups defined by year of diagnosis or age at diagnosis. When estimating expected survival we require the expected probabilities of death according to age and year at time of follow-up (rather than time of diagnosis). The command must therefore keep track of both. We have adopted the convention of prefixing variable names with an underscore when they are updated with follow-up, for example, the variable `age` carries age at diagnosis and `_age` carries attained age. By default, the patient data file should contain variables named `age` and `yydx` but cannot contain variables named `_age` and `_year`. The popmort file, on the other hand, should contain variables `_age` and `_year` since the expected probabilities are merged using these 'time-updated' variables. Alternative variable names can be specified using the appropriate option.

3.4 Example 1 – life table estimates of relative survival

We will illustrate the commands using data provided by the Finnish Cancer Registry on patients diagnosed with colon carcinoma in Finland 1975–1994. These data are distributed with the package along with do files to reproduce all analyses presented in this paper. We first estimate life tables for each gender (only the table for males is shown) among patient with clinically localised (`stage==1`) disease. We have chosen to use six-month intervals for the first two intervals followed by annual intervals up to 10 years.

```
. use colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. gen id=_n
. stset surv_mm, fail(status==1 2) id(id) scale(12)
  (output omitted)
. strs using popmort if stage==1, br(0 0.5 1(1)10) mergeby(_year sex _age) by(s
> ex) list(start end n d w cp cp_e2 cr_e2)
         failure _d:  status == 1 2
   analysis time _t:  surv_mm/12
                 id:  id
No late entry detected - p is estimated using the actuarial method
```

```
-> sex = Male
```

start	end	n	d	w	cp	cp_e2	cr_e2
0	.5	2620	229	0	0.9126	0.9728	0.9381
.5	1	2391	99	0	0.8748	0.9484	0.9224
1	2	2292	229	166	0.7841	0.8993	0.8719
2	3	1897	180	139	0.7069	0.8517	0.8300
3	4	1578	140	119	0.6417	0.8048	0.7974
4	5	1319	113	104	0.5845	0.7588	0.7703
5	6	1102	102	81	0.5283	0.7143	0.7396
6	7	919	71	71	0.4859	0.6721	0.7229
7	8	777	59	72	0.4472	0.6312	0.7084
8	9	646	49	62	0.4115	0.5921	0.6950
9	10	535	33	58	0.3847	0.5545	0.6937

Columns in the life table are number first at risk (`n`), deaths (`d`), censorings (`w`), cumulative observed survival (`cp`), Ederer II cumulative expected survival (`cp_e2`), and cumulative relative survival (`cr_e2`). The estimated 1-year relative survival ratio is 0.922 and the estimated 5-year relative survival ratio is 0.770. Other quantities provided by default but omitted here due to space limitations are interval-specific observed survival (`p`), interval-specific expected survival (`p_star`), interval-specific relative survival (`r`) and 95% confidence intervals for the cumulative relative survival ratio.

When we `stset` the data all deaths are classified as events (values 1 and 2 of the variable `status` indicate death due to cancer and non-cancer respectively). The data did not initially contain an id variable so we were required to create one (a requirement of the `stsplit` command called by `strs`). We made use of the variable `surv_mm` (containing time from diagnosis to death or censoring in months) to `stset` the data. The timescale must be time since entry in years so we have applied a scale factor of 12. Variables containing dates of diagnosis (`dx`) and exit (`exit`) could have also been used to `stset` the data (see the next example).

Because the life table estimates can be saved to a Stata data set (see the `save` option) it is simple to produce graphs or tables of quantities of interest. For example, we can tabulate the number of patients initially at risk along with the 5-year observed and relative survival for each combination of age and sex.

```
. strs using popmort if stage==1, br(0(1)10) mergeby(_year sex _age) by(sex age
> grp) save(replace)
  (output omitted)
. use grouped, clear
(Collapsed (or grouped) survival data)
. gen n0=n[_n-4]
(4 missing values generated)
. list sex agegrp n0 cp cr_e2 lo_cr_e2 hi_cr_e2 if end==5, sepby(sex) noobs
```

sex	agegrp	n0	cp	cr_e2	lo_cr_e2	hi_cr_e2
Male	0-44	161	0.7737	0.7881	0.7102	0.8486
Male	45-59	462	0.7686	0.8233	0.7766	0.8636
Male	60-74	1228	0.5945	0.7512	0.7128	0.7878
Male	75+	769	0.4131	0.7777	0.7067	0.8479
Female	0-44	136	0.7657	0.7709	0.6866	0.8358
Female	45-59	531	0.7765	0.7953	0.7536	0.8314
Female	60-74	1488	0.6993	0.7873	0.7588	0.8141
Female	75+	1499	0.4854	0.7816	0.7374	0.8249

We see that the overall 5-year survival (`cp`) decreases with age as expected but 5-year relative survival (`cr`) is similar across categories of age and sex. We could also use the data in `grouped.dta` to, for example, plot survival estimates as a function of follow-up time.

3.5 Example 2 – expected survival using three different methods

A description of the three different methods for estimating expected survival is given in Section 2.2. To obtain estimates of expected survival using the Hakulinen method we must specify, using the `potfu()` option, a variable containing the last date of potential follow-up for each patient. The `ederer1` option results in Ederer I estimates of expected and relative survival also being estimated. Ederer II estimates are produced by default and no option is required.

```
. use colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. gen id=_n
. stset exit, origin(dx) fail(status==1 2) id(id) scale(365.24)
```

(*output omitted*)

```
. gen long potfu = date("31/12/1995","dmy")
. strs using popmort if stage==1, br(0(1)10) mergeby(_year sex _age) by(sex) li
> st(start n d w cr_e1 cr_e2 cr_hak) ederer1 potfu(potfu)
        failure _d:  status == 1 2
  analysis time _t:  (exit-origin)/365.24
            origin:  time dx
                id:  id
No late entry detected - p is estimated using the actuarial method
```

```
-> sex = Male
```

start	end	n	d	w	cr_e1	cr_e2	cr_hak
0	1	2620	328	0	0.9238	0.9238	0.9238
1	2	2292	229	166	0.8758	0.8732	0.8756
2	3	1897	180	139	0.8361	0.8312	0.8359
3	4	1578	140	119	0.8050	0.7986	0.8049
4	5	1319	113	104	0.7787	0.7715	0.7787
5	6	1102	102	81	0.7486	0.7407	0.7487
6	7	919	71	71	0.7333	0.7239	0.7335
7	8	777	59	72	0.7200	0.7095	0.7202
8	9	646	49	62	0.7082	0.6961	0.7082
9	10	535	33	58	0.7085	0.6948	0.7087

(*output omitted*)

We see only small differences between the estimates of cumulative relative survival made using the Ederer I (`cr_e1`), Ederer II (`cr_e2`), and Hakulinen (`cr_hak`) methods. Differences between the three methods are, in general, small during the first 10 years of follow-up particularly. The Ederer II and Hakulinen estimates are generally similar if analyses are stratified by age since such stratification reduces any existing heterogeneity in withdrawal patterns.

3.6 Example 3 - cohort, complete, period, and hybrid estimation

The primary purpose of this section is to demonstrate how period and hybrid estimates of relative survival can be obtained using `strs`. We will estimate 10-year survival of patients diagnosed with localised (`stage==1`) colon carcinoma in Finland. Our data set includes all patients diagnosed 1975–1994 with follow-up until the end of 1995. We adopt the same terminology for the various approaches (cohort, complete, period, hybrid) employed by (Brenner et al. 2004b) although note that this terminology is not used consistently in the literature. The fundamental difference between the various approaches is in the definition of person-time at risk that contributes to the analysis. As such, the call to `strs` is similar for each approach.

Cohort approach

To estimate 10-year survival using a cohort approach, all patients must have a potential follow-up of at least 10 years. Our data set includes patients diagnosed 1975–1994 with follow-up until the end of 1995. Therefore, only patients diagnosed 1985 or earlier can contribute to the cohort estimate of 10-year survival. This is easily implemented in Stata.

```
. strs using popmort if stage==1 & yydx <= 1985, br(0(1)10) ///
                        mergeby(_year sex _age) by(sex)
```

Such estimates, based on patients diagnosed at least 10 years in the past, will clearly not be relevant for recently diagnosed patients.

Complete approach

Before the introduction of period analysis, up-to-date estimates of patient survival were typically made using the so-called complete approach. To estimate 10-year survival we must include some patient diagnosed more than 10 years ago but we also include recently diagnosed patients, even though they cannot be followed for 10 years. The cumulative 10-year survival is estimated as a product of conditional survival probabilities where the recently diagnosed patients contribute to only some of the conditional estimates. We would therefore include patients diagnosed up until 1994 (i.e., as recent as possible) but must, at a minimum, include patients diagnosed as far back as 1985. In order to improve precision without overly sacrificing recency, we might decide to also include patients diagnosed in 1994. That is, the conditional survival probability for the 10th year will be based on those patients diagnosed in 1984 and 1985 who survived at least 9 years.

```
. strs using popmort if stage==1 & yydx >= 1984, br(0(1)10) ///
                        mergeby(_year sex _age) by(sex)
```

Although more up-to-date than cohort estimates, these estimates are still heavily influenced by the survival experience of patients diagnosed many years in the past.

Period approach

To overcome this drawback, Brenner and colleagues suggested that lifetable estimates of patient survival could be made using a period rather than cohort approach (Brenner et al. 2004b; Brenner and Gefeller 1996). Time at risk is left truncated at the start of the period window and right censored at the end. If we consider the previous example using the complete approach, the conditional survival for the first year is based on patients diagnosed during an 11 year period (1984–1994) and conditional survival for the second year is based on patients diagnosed during a 10 year period (1984–1993). With period analysis, each conditional probability is estimated based only the survival experience of only recently diagnosed patients. There is a trade-off between precision and recency; a narrow period window (e.g., 1 year) will improve recency but reduce precision compared to a wider (e.g., 5 year) period window.

Period analysis has been shown to provide more accurate predictions of the prognosis of newly diagnosed patients and is able to detect temporal trends in patient survival sooner than the traditional cohort approach (Brenner et al. 2004b). Our approach to period estimation using Stata is to first identify the time at risk during the period window for each individual by applying `stset` with calendar time as the timescale. For example, we might be interested in the period between 1 January 1990 and 31 December 1994 (the last five years for which incidence data were collected in this dataset).

```
. stset exit, origin(dx) enter(time mdy(1,1,1990)) exit(time mdy(12,31,1994)) ///
        f(status==1 2) id(id) scale(365.24)
```

We can then apply `strs` in the usual manner to obtain Ederer II estimates

```
. strs using popmort if stage==1, br(0(1)10) mergeby(_year sex _age) by(sex)
```

or Hakulinen estimates

```
. replace potfu = date("31/12/1994","dmy")
. strs using popmort if stage==1, br(0(1)10) potfu(potfu) by(sex) ///
         mergeby(_year sex _age)
```

Note that if an individual dies before the start of the period window the record is marked with `_st=0` and is not considered in analyses performed using `st` commands. Although such individuals do not contribute to the estimates of observed survival, they do contribute to the estimation of expected survival using the Hakulinen method.

Hybrid approach

Application of the period approach may be problematic if the follow-up period extends beyond the period for which incident cases are accrued. For example, our sample data set contains patients diagnosed up until December 1994 with follow-up until December 1995. For this reason, we censored the follow-up of all individuals on 31st December 1994 in the previous example.

What would we do if we wanted to perform period analysis with a window from 1

January 1991 – 31 December 1995? Using annual intervals, the first conditional estimate would contain contributions from patients diagnosed 1990–1994, the second would contain contributions from patients diagnosed 1989–1994, and the third conditional estimate would contain contributions from patients diagnosed 1988–1993. All conditional estimates contain contributions from 6 potential years of diagnosis, apart from the first year which only contains contributions from 5 potential years of diagnosis. Brenner and Rachet (2004) suggested that, in such a situation, the period window should be widened for the first year (it should be made 1 January 1990 – 31 December 1995 so patient diagnosed 1989–1994 will contribute person-time). They called this approach the 'hybrid approach'. The distinctive feature of the hybrid approach is that the date at which individuals become at risk (the start of the period window) differs according to year of diagnosis. This is relatively easy to apply in Stata:

```
. gen long hybridtime = cond(yydx>1989, dx, mdy(1,1,1991))
. stset exit, origin(dx) enter(time hybridtime) f(status==1 2) id(id) scale(365.24)
. replace potfu = date("31/12/1995","dmy")
. strs using popmort if stage==1, br(0(1)10) potfu(potfu) by(sex) ///
        mergeby(_year sex _age)
```

We create a new variable `hybridtime` to hold the date at which each individual becomes at risk. This corresponds to the date of diagnosis for patients diagnosed 1990–1994 and to 1 January 1991 for patients diagnosed before 1 January 1990. A diagram such as the one used in Brenner and Rachet (2004) can assist in defining the entry dates. We then `stset` the data with this as the start of the time at risk (using the `enter()` option) and call `strs` in the usual manner. Table 1 shows 10-year relative survival estimates (Hakulinen method) for patients diagnosed with colon carcinoma according to the four different approaches.

Approach	RSR_{males}	$RSR_{females}$
Cohort	0.6831	0.7050
Complete	0.7002	0.7358
Period	0.7094	0.7880
Hybrid	0.7415	0.7840

Table 1: 10-year relative survival for patients diagnosed with localised colon carcinoma in Finland 1985-1994 using four different approaches

3.7 Example 4 – age-standardised relative survival estimates

In this section we will discuss age-standardisation although one may standardise on factors other than age. Age-standardisation can be employed to facilitate comparisons of relative survival between different populations, such as patients diagnosed in different calendar periods. Although relative survival estimates are automatically adjusted for differences in expected survival due to differing age distributions, they are not adjusted

age (i)	n_i	RSR_i	w_i
0-44	381	0.4458	0.042
45-59	1339	0.4912	0.147
60-74	3699	0.4546	0.407
75+	3668	0.3871	0.404
Crude	9087	0.4358	
Age-standardised		0.4324	

Table 2: Age-specific numbers of patients (n_i) and estimates of 10-year relative survival (RSR_i) for patients diagnosed with colon carcinoma in Finland 1985–1994

for the fact that relative survival (excess mortality) may also depend on age.

Hakulinen (1977) suggested that one should consider using age standardisation even when estimating relative survival for a single population where there is no interest in making comparisons. He showed that it is possible for the age-specific survivor function to be constant after a certain follow-up time (indicating no excess mortality) in each and every age stratum but for the all-age survivor function to increase. This situation arises because the cumulative survival is the product of conditional survival proportions, each with a different age distribution. Professor Hakulinen considered it counterintuitive that the 'all ages' curve should have a different shape to the common shape of the age-specific curves and suggested that all-age estimates be age standardized (using the age distribution at the start of follow-up as the standard population). This is traditional direct standardisation using an internal standard. Table 2 shows crude and age-specific estimates of 10-year survival for patients diagnosed with colon carcinoma in Finland 1985–1994.

If we directly age-standardise using the traditional method with an internal standard the weights (w_i) are simply the proportion of patients in each age group at the start of follow-up. The age-standardised 10-year RSR is given by $\sum_i RSR_i w_i / \sum_i w_i = 0.4324$. Specifying the `standstrata()` option results in `strs` first producing stratified life tables for each level of the variables specified in `standstrata()` and then producing standardised estimates using the weights contained in the variable specified in the `iweights()` option.

```
. stset exit, origin(dx) f(status==1 2) id(id) scale(365.24)
. recode agegrp 0=0.041928 1=0.147353 2=0.407065 3=0.403654, gen(standwei)
. strs using popmort [iw=standwei] if yydx > 1984, br(0(1)20) ///
        mergeby(_year sex _age) standstrata(agegrp) notables
```

(*output omitted*)

The weights should be specified as proportions. In this example, the crude and internally age-standardised estimates were similar although this is not always the case (Hakulinen 1977). It is possible to use the `by()` option together with `standstrata()` in order to produce, for example, age-standardised estimates for each calendar period. For example, the following code produces age-standardised estimates for each period using the age structure for the latter period as the standard. The variable `year8594` is an indicator

	10-year relative Survival		
Period	Crude	Age-standardised (traditional)	Age-standardised (alternative)
1975–1984	0.4035	0.4023	0.3998
1985–1994	0.4358	0.4324	0.4358

Table 3: Crude, age-standardised and age adjusted (alternative) estimates of 10-year relative survival obtained in each period for patients with colon carcinoma in Finland. The age distribution for 1985–1994 is used as the standard population.

for diagnosis during the period 1985–1994 (versus 1975–1984).

```
. stset exit, origin(dx) f(status==1 2) id(id) scale(365.24)
. recode agegrp 0=0.041928 1=0.147353 2=0.407065 3=0.403654, gen(standwei)
. strs using popmort [iw=standwei], br(0(1)20) mergeby(_year sex _age) ///
        standstrata(agegrp) by(year8594) notables
```

(*output omitted*)

Rather than weighting based on the age distribution at the start, Brenner et al. (2004a) suggest using a weights that changes throughout follow-up time. This is achieved by assigning individual weights to each patient and constructing a weighted life table (Brenner et al. 2004a). Specifying the `brenner` option causes `strs` to produce standardised estimates using this 'alternative' method. A property of this method is that if we use the actual age distribution of the patients as the standard population then the age-standardised estimates will, unlike the traditional method, be identical to the crude estimates (see table 3). Table 3 shows crude, age-standardised and age-adjusted (alternative) 10-year relative survival estimates for each period. The two groups under comparison have a very similar age structure so there are only small differences between the different approaches although this is not always the case (Brenner et al. 2004a). The same technique can be used with respect to other factors, such as race or stage, but modelling is generally the method of choice for comparing survival between populations after adjustment for multiple covariates.

4 Modelling excess mortality

The mortality analogue of relative survival is excess mortality and it is this quantity that is modelled. The total hazard 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,

$$\lambda(t;\mathbf{z}) = \lambda^*(t;\mathbf{z}) + \nu(t;\mathbf{z}). \tag{1}$$

The expected hazard is annotated with an asterisk to indicate that it is estimated from external data (general-population mortality rates). 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 $\mathbf{z}$.

Follow-up time is partitioned into bands corresponding to life table intervals. These are typically of length one year although it is possible to use shorter intervals early in the follow up where mortality is often higher and changing rapidly (as in Section 3.4). A set of indicator variables are constructed (one indicator variable for each interval excluding the reference interval) and incorporated into the covariate matrix. 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

$$\lambda(\mathbf{x}) = \lambda^*(\mathbf{x}) + \exp(\mathbf{x}\beta). \tag{2}$$

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 2 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 (Suissa 1999). An excess hazard ratio of, for example, 1.5 for males compared to females implies that the excess mortality associated with a diagnosis of cancer is 50% higher for males than females.

4.1 Modelling excess mortality using a full likelihood approach

Estève et al. (1990) described a method for estimating the model in Equation 2 directly from individual-level data using a full maximum likelihood approach. The likelihood function is

$$L = \prod_{i=1}^{n} \exp(-\int_0^{t_i} \lambda(s)\, ds)[\lambda(t_i)]^{d_i}, \tag{3}$$

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

$$l(\beta) = -\sum_{i=1}^{n} \int_0^{t_i} \lambda^*(s)\, ds - \sum_{i=1}^{n} \int_0^{t_i} \nu(s)\, ds + \sum_{i=1}^{n} d_i \ln[\lambda^*(t_i) + \nu(t_i)]. \tag{4}$$

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

$$l(\beta) = \sum_{j=1}^{J} \left[d_j \ln[\lambda^*(\mathbf{x}_j) + \exp(\mathbf{x}_j\beta)] - y_j \exp(\mathbf{x}_j\beta) \right]. \tag{5}$$

We can use the `ml` command to maximise the log likelihood function shown in Equation 5. The likelihood is defined in `esteve.ado`.

```
program define esteve
version 7
args lnf theta
qui replace `lnf'=-exp(`theta')*y if $ML_y1==0
qui replace `lnf'=ln(-ln(p_star)+exp(`theta'))-exp(`theta')*y if $ML_y1==1
end
```

Example

We fit the model to the colon carcinoma data restricting the analysis to the first five years of follow-up. After declaring the data to be survival time (using `stset`) we call `strs` to tabulate the numbers of observed and expected deaths for each combination of follow-up interval, sex, calendar period, and age group. We have suppressed the display of life tables (the `notables` option) but have requested estimates be saved using the default file names (`individ.dta` and `grouped.dta`). The full likelihood model is estimated using `individ.dta` (which contains one observation for each individual for each life table interval).

```
. use colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. gen id=_n
. stset surv_mm, fail(status==1 2) id(id) scale(12)
. strs using popmort if stage==1, br(0(1)5) mergeby(_year sex _age) ///
    by(sex year8594 agegrp) save(replace) notable
. use individ, clear
(Survival data containing individual subject-band observations)
. xi: ml model lf esteve (d=i.end i.sex i.year8594 i.agegrp)
i.end             _Iend_1-5           (naturally coded; _Iend_1 omitted)
i.sex             _Isex_1-2           (naturally coded; _Isex_1 omitted)
i.year8594        _Iyear8594_0-1      (naturally coded; _Iyear8594_0 omitted)
i.agegrp          _Iagegrp_0-3        (naturally coded; _Iagegrp_0 omitted)
. ml maximize, eform("RER")
                                                Number of obs   =      23579
                                                Wald chi2(9)    =      72.73
Log likelihood = -5969.5775                     Prob > chi2     =     0.0000
------------------------------------------------------------------------------
```

d	RER	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iend_2	.8286045	.0779917	-2.00	0.046	.689015	.9964739
_Iend_3	.6765733	.0727639	-3.63	0.000	.5479868	.835333
_Iend_4	.5383155	.069149	-4.82	0.000	.4185008	.6924325
_Iend_5	.4606403	.0690407	-5.17	0.000	.343387	.617931
_Isex_2	.9545966	.0737863	-0.60	0.548	.8203999	1.110744
_Iyear8594_1	.734979	.055002	-4.11	0.000	.6347102	.8510879
_Iagegrp_1	.8663227	.135108	-0.92	0.358	.6381604	1.17606
_Iagegrp_2	1.055003	.1508525	0.37	0.708	.7971545	1.396256
_Iagegrp_3	1.341785	.2022822	1.95	0.051	.9985251	1.803045

The estimates are identical to those presented in Table I of Dickman et al. (2004). The variable `year8594` is coded as 1 for patients diagnosed 1985–1994 and 0 for patients diagnosed 1975–1984. We see that patients diagnosed in the recent period are estimated to experience 27% lower excess mortality compared to those diagnosed in the earlier period. There is evidence that excess mortality decreases with follow-up time, some evidence of higher excess mortality in the oldest age group, and no evidence of a difference between males and females.

4.2 Modelling excess mortality using Poisson regression

The relative survival model (Equation 2) assumes piecewise constant hazards which implies a Poisson process for the number of deaths in each interval. 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. 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. Equation 2 is then written as

$$\mu_j/y_j = d_j^*/y_j + \exp(\mathbf{x}\beta), \tag{6}$$

which can be written as

$$\ln(\mu_j - d_j^*) = \ln(y_j) + \mathbf{x}\beta, \tag{7}$$

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 $\ln(y_j)$. This is not a standard link function so is defined in `rs.ado`.

```
program define rs
version 7
args todo eta mu return
if ‘todo’ == -1
         global SGLM_lt "Relative survival"
         global SGLM_lf "log(u-d*)"
                 exit

        if ‘todo’ == 0
                 gen double ‘eta’ = ln(‘mu’-$SGLM_p)
```

```
                exit

        if `todo' == 1
                gen double `mu' = exp(`eta')+$SGLM_p
                exit

        if `todo' == 2
                gen double `return' = exp(`eta')
                exit

        if `todo' == 3
                gen double `return' = exp(`eta')
                exit

        di as error "Unknown call to glm link function"
        exit 198
end
```

Example: Poisson regression

The `strs` command in the previous example produced two output data files, `individ.dta` containing one observation for each subject-band and `grouped.dta` containing one observation for each life table interval. We will fit the Poisson regression model to the grouped data; if we fitted the model to the data in `individ.dta` we would obtain identical estimates to the full likelihood approach (Section 4.1) since we would be maximising the same likelihood using the same data.

```
. use grouped, clear
(Collapsed (or grouped) survival data)
. xi: glm d i.end i.sex i.year8594 i.agegrp, fam(pois) link(rs d_star) lnoffset
> (y) eform
Generalized linear models                         No. of obs      =        80
Optimization     : ML                             Residual df     =        70
                                                  Scale parameter =         1
Deviance         =  131.4342128                   (1/df) Deviance =  1.877632
Pearson          =  130.1530694                   (1/df) Pearson  =   1.85933
Variance function: V(u) = u                       [Poisson]
Link function    : g(u) = log(u-d*)               [Relative survival]
                                                  AIC             =   6.39959
Log likelihood   = -245.9836017                   BIC             = -175.3077
```

d	ExpB	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
_Iend_2	.7984084	.0730515	-2.46	0.014	.6673339	.955228
_Iend_3	.6230213	.0671961	-4.39	0.000	.5043086	.7696785
_Iend_4	.4969433	.0645561	-5.38	0.000	.3852391	.6410374
_Iend_5	.4334347	.065147	-5.56	0.000	.322838	.5819191
_Isex_2	.9564493	.0729823	-0.58	0.560	.8235891	1.110742
_Iyear8594_1	.7308044	.0539291	-4.25	0.000	.6323935	.8445296
_Iagegrp_1	.8642841	.1353083	-0.93	0.352	.635911	1.174672
_Iagegrp_2	1.071568	.1534869	0.48	0.629	.8092774	1.418869
_Iagegrp_3	1.436319	.2146593	2.42	0.015	1.071613	1.925147
y	(exposure)					

This model is conceptually identical to the full likelihood approach applied in the previous section and the estimates are very similar. The advantage of estimating the model in the framework of generalised linear models is that we have access to a rich theoretical framework and can utilise, for example, regression diagnostics. An advantage of fitting the model to collapsed data is that we can assess goodness-of-fit using the deviance Pearson chi square statistics (provided the data are non-spare). We see that there is evidence of lack of fit (deviance is 131.4 with 70 df) and further investigation reveals that an age by follow-up interaction is required (see Dickman et al. 2004, Table II).

Example: Poisson regression using smoothing splines

We have assumed the hazard is piecewise constant (i.e., a step function) over follow-up time, an assumption that is not attractive from a clinical/biological perspective. We might alternatively specify narrower time-bands (e.g., monthly) and model the effect of follow-up using a natural cubic spline.

```
. strs using popmort if stage==1, br(0(0.083333333333)5) mergeby(_year sex _age)
> by(sex year8594 agegrp) save(replace)
. use grouped, clear
. spbase end, gen(endb)
. xi: glm d $endb i.sex i.year8594 i.agegrp, fam(pois) link(rs d_star) lnoff(y) ef
```

The same approach can be used for any metric variable, for example, age at diagnosis. Alternative methods for fitting smooth functions, such as fractional polynomials (Lambert et al. 2005), restricted cubic splines (using the `rc_spline` command), or B-splines (Giorgi et al. 2003) can also be applied.

As an illustration of assessing the goodness-of-fit of this model, figure 1 shows the model-based estimates of relative survival for each age group for males with localised colon cancer diagnosed in 1985-1994 and corresponding empirical estimates with 95% confidence intervals.

```
. predict xb, xb nooffset                          // excess risk
. gen r_hat = exp(-exp(xb)*0.083333333)            // interval relative survival
. bysort sex year8594 agegrp (end) : ///
            g rs_hat = exp(sum(log(r_hat)))        // cumulative relative survival
. twoway (rcap  lo_cr_h hi_cr_h end if end==int(end) & sex==1 & year8594==1) ///
      (scatter cr_hak end if end==int(end) & sex==1 & year8594==1) ///
      (line rs_hat end if sex==1 & year8594==1, lw(medthick)), ///
      by(agegrp, legend(off)) yti("Relative Survival") ///
      xti("Years from diagnosis") xla(0(1)5) yla(0.6(.1)1)
```

As is often the case with cancer survival data, patients aged 75 years of more at diagnosis have considerably higher mortality during the first year following diagnosis but once they have survived the first year experience excess mortality more similar to the other age groups. That is, the excess hazards non-proportional by age at diagnosis.

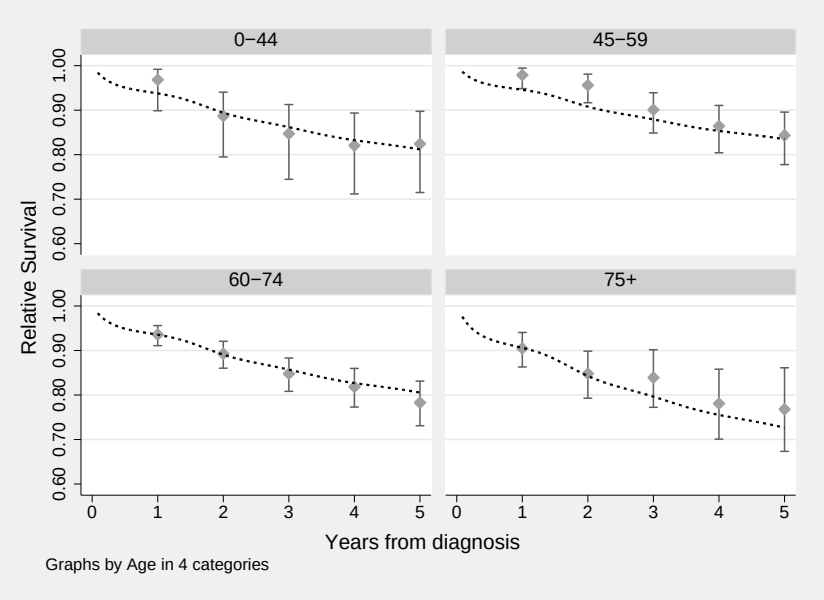

Figure 1: Model-based (dotted line) and empirical (with 95% CI) estimates of relative survival by age groups for males with localised colon cancer diagnosed in 1985-1994

4.3 Hakulinen-Tenkanen approach to modelling excess mortality

Grouped survival data can be modelled in the framework of generalised linear models by assuming the number of patients surviving the interval follows a binomial distribution with denominator the effective number at risk and using a complementary log-log link. Hakulinen and Tenkanen (1987) extended this approach to relative survival where the link function is now complementary log-log combined with a division by the expected survival proportion p_j^*. That is,

$$\ln\left[-\ln\frac{p_j}{p_j^*}\right] = \mathbf{x}\beta. \tag{8}$$

We note that $-\ln(p_j/p_j^*)$ is the cumulative excess hazard for interval j so this approach, as with the two previous approaches, equates the natural logarithm of the excess hazard with the linear predictor. This link function is not standard so, as with the Poisson regression model for excess mortality, the link function is defined in an ado file (`ht.ado`) and the model estimated using the `glm` command in the usual manner.

```
use grouped, clear
xi: glm ns i.end i.sex i.year8594 i.agegrp, fam(bin n_prime) link(ht p_star)
  (output omitted)
```

5 Appendix

Formulae for estimation of expected survival

Under the Ederer I method (Ederer et al. 1961), the cumulative expected survival from the date of diagnosis to the end of the ith interval is given by

$$ {}_1p_i^* = \sum_{h=1}^{l_1} {}_1p_i^*(h)/l_1, $$

where l_1 is the total number of patients alive at the start of follow-up and ${}_1p_i^*(h)$ is the expected probability of surviving to the end of the ith interval for a person in the general population, similar to the hth patient alive at the beginning of follow-up with respect to age, sex, and calendar time, given by

$$ {}_1p_i^*(h) = \prod_{j=1}^{i} p_j^*(h). $$

Under the Ederer II method (Ederer and Heise 1959)

$$ {}_1p_i^* = \prod_{j=1}^{i} p_{j2}^*, $$

where

$$ p_{j2}^* = \sum_{h=1}^{l_j} p_j^*(h)/l_j $$

is the average of the annual expected survival probabilities $p_j^*(h)$ of the patients alive at the start of the jth interval.

The expected survival proportion using the Hakulinen method (Hakulinen 1982) is derived as follows. Let k_j be the number of patients with a potential follow-up time which extends beyond the beginning of the jth interval. Let the first k_{ja} of these k_j patients have a potential follow-up time which extends past the end of the jth interval and the last k_{jb} be potential withdrawals during the jth interval. It follows that $k_1 = l_1$, $k_{j+1} = k_{ja}$, and $k_j = k_{ja} + k_{jb}$. We will use the notation K_{ja} to refer to the set of k_{ja} patients etc. and h to index the k_{ja} patients in the set K_{ja}. The expected number of patients alive and under observation at the beginning of the jth interval is given by:

$$ l_j^* = \begin{cases} \sum_{h \in K_j} {}_1p_{j-1}^*(h) & \text{for } j \geq 2 \\ l_1 & \text{for } j = 1 \end{cases} $$

For the k_{jb} patients with potential follow-up times ending during the jth interval, it is assumed that each patient is at risk for half of the interval, so the expected probability of dying during the interval is given by $1 - \sqrt{p_j^*}$. The expected number of patients

withdrawing alive during the jth interval is therefore given by:

$$w_j^* = \begin{cases} \sum_{h \in K_{jb}} {}_1p_{j-1}^*(h)\sqrt{p_j^*(h)} & \text{for } j \geq 2 \\ \sum_{h \in K_{1b}} \sqrt{p_1^*(h)} & \text{for } j = 1 \end{cases}$$

The expected number of patients dying during the jth interval, among the k_{jb} patients with potential follow-up time ending during the same interval is given by:

$$\delta_j^* = \begin{cases} \sum_{h \in K_{jb}} {}_1p_{j-1}^*(h)[1 - \sqrt{p_j^*(h)}] & \text{for } j \geq 2 \\ \sum_{h \in K_{1b}} [1 - \sqrt{p_1^*(h)}] & \text{for } j = 1 \end{cases}$$

and the expected total number of patients dying during the jth interval is given by:

$$d_j^* = \begin{cases} \left\{\sum_{h \in K_{ja}} {}_1p_{j-1}^*(h)[1 - p_j^*(h)]\right\} + \delta_j^* & \text{for } j \geq 2 \\ \left\{\sum_{h \in K_{1a}} [1 - p_1^*(h)]\right\} + \delta_1^* & \text{for } j = 1 \end{cases}$$

The expected interval-specific survival proportion is then written as:

$$g_j^* = 1 - d_j^*/(l_j^* - w_j^*/2),$$

and, finally, the expected survival proportion from the beginning of follow-up (usually diagnosis) to the end of the ith interval is obtained by calculating:

$${}_1p_i^* = \prod_{j=1}^{i} g_j^*.$$

6 Acknowledgements

We thank Andy Sloggett for his contribution to discussions surrounding the underlying methodology and Stata programming and the Finnish Cancer Registry for providing data. Paul Dickman thanks Cancerfonden for financial support.

7 References

Brenner, H., V. Arndt, O. Gefeller, and T. Hakulinen. 2004a. An alternative approach to age adjustment of cancer survival rates. *Eur J Cancer* 40(15): 2317–2322. *http://dx.doi.org/10.1016/j.ejca.2004.07.007*

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

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

Brenner, H. and T. Hakulinen. 2005. Substantial overestimation of standard errors of relative survival rates of cancer patients. *Am J Epidemiol* 161(8): 781–786. *http://dx.doi.org/10.1093/aje/kwi099*

Brenner, H. and B. Rachet. 2004. Hybrid analysis for up-to-date long-term survival rates in cancer registries with delayed recording of incident cases. *European Journal of Cancer* 40: 2494–501.

Coleman, M. P., P. Babb, P. Damiecki, P. Grosclaude, S. Honjo, J. Jones, G. Knerer, A. Pitard, M. Quinn, A. Sloggett, and B. De Stavola. 1999. *Cancer Survival Trends in England and Wales, 1971–1995: Deprivation and NHS Region*. No. 61 in Studies in Medical and Population Subjects, London: The Stationery Office.

Dickman, P. W. and H.-O. Adami. 2006. Interpreting trends in cancer patient survival. *Journal of Internal Medicine* 260: 103–17.

Dickman, P. W., A. Sloggett, M. Hills, and T. Hakulinen. 2004. Regression models for relative survival. *Stat Med* 23(1): 51–64. *http://dx.doi.org/10.1002/sim.1597*

Ederer, F., L. M. Axtell, and S. J. Cutler. 1961. The Relative Survival Rate: A Statistical Methodology. *National Cancer Institute Monograph* 6: 101–121.

Ederer, F. and H. Heise. 1959. Instructions to IBM 650 Programmers in Processing Survival Computations. Methodological note No. 10, End Results Evaluation Section, National Cancer Institute, Bethesda MD.

Estève, J., E. Benhamou, M. Croasdale, and L. Raymond. 1990. Relative Survival and the Estimation of Net Survival: Elements for Further Discussion. *Statistics in Medicine* 9: 529–538.

Gamel, J. W. and R. L. Vogel. 2001. Non-parametric comparison of relative versus cause-specific survival in Surveillance, Epidemiology and End Results (SEER) programme breast cancer patients. *Statistical Methods in Medical Research* 10(5): 339–352.

Giorgi, R., M. Abrahamowicz, C. Quantin, P. Bolard, J. Esteve, J. Gouvernet, and J. Faivre. 2003. A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat Med* 22(17): 2767–84. *http://dx.doi.org/10.1002/sim.1484*

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

Hakulinen, T. 1977. On Long-Term Relative Survival Rates. *Journal of Chronic Diseases* 30: 431–443.

—. 1982. Cancer Survival Corrected for Heterogeneity in Patient Withdrawal. *Biometrics* 38: 933–942.

Hakulinen, T. and L. Tenkanen. 1987. Regression Analysis of Relative Survival Rates. *Applied Statistics* 36: 309–317.

Lambert, P. C., L. K. Smith, D. R. Jones, and J. L. Botha. 2005. Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects. *Statistics in Medicine* 24: 3871–85.

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