The Stata Journal (2015)
15, Number 1, pp. 186–215

Estimating and modeling relative survival

Paul W. Dickman
Karolinska Institutet
Stockholm, Sweden
paul.dickman@ki.se

Enzo Coviello
ASL BT
Barletta, Italy
enzo.coviello@tin.it

Abstract. When estimating patient survival using data collected by population-based cancer registries, it is common to estimate net survival in a relative-survival framework. Net survival can be estimated using the relative-survival ratio, which is 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. In this article, we describe a command, `strs`, for life-table estimation of relative survival. We discuss three methods for estimating expected survival, as well as the cohort, period, and hybrid approaches for estimating relative survival. We also implement a life-table version of the Pohar Perme (2012, *Biometrics* 68: 113–120) estimator of net survival, and we describe two methods for age standardization. We also explain how, in addition to net probabilities of death, crude probabilities of death due to cancer and due to other causes can be estimated using the method of Cronin and Feuer (2000, *Statistics in Medicine* 19: 1729–1740). We conclude this article with discussion and examples of modeling excess mortality using various approaches, including the full-likelihood approach (using the `ml` command) and Poisson regression (using the `glm` command with a user-specified link function).

Keywords: st0376, strs, excess mortality, relative survival, survival analysis, Poisson regression, life table, cancer survival, period analysis

1 Introduction

When studying population-based cancer survival (that is, the estimation of patient survival using data collected by population-based cancer registries), we are typically interested in estimating the probability that patients will die of their specific cancer. That is, we have a competing-risks setting and can choose to estimate either net or crude probabilities of death due to cancer (Tsiatis 2005). A common approach with competing-risks data is to classify the cause of death of each individual who dies and use this information to estimate what is commonly called cause-specific survival. Such an approach can be problematic with cancer registry data, because information on cause of death is often unreliable or unavailable (Gamel and Vogel 2001). As such, it is common in population-based cancer survival to instead estimate the chosen measure (crude or net probability) in a relative-survival (RS) framework, where cause-of-death information is not required (Dickman and Adami 2006; Dickman et al. 2004; Estève et al. 1990).

In an RS framework, we estimate the excess mortality rate as the difference between the total (all-cause) mortality rate among the patients and the expected mortality rate 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. The major advantages of using an RS framework are that it does not require information on cause of death and that it provides a measure of the excess mortality experienced by patients diagnosed with cancer, regardless of whether the excess mortality is directly or indirectly (for example, due to treatment complications) attributable to the cancer. RS, the survival analogue of excess mortality, 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 common to estimate expected survival from nationwide population life tables stratified by age, sex, calendar time, and, where applicable, race.

In this article, we describe how to estimate various measures, such as crude and net probabilities of death, in an RS framework, and we explain how to implement these measures in Stata. Our focus is on population-based cancer survival, although the utility of the methodologic approach is not restricted to studying cancer (Nelson et al. 2008).

2 Overview of approaches

It is important to distinguish between the measures (crude and net probabilities), the framework (cause-specific or relative) for estimating the chosen measure, and the estimators available within the chosen framework.

- Crude survival (measure)
  - Cause-specific framework
    - Standard estimators of the cumulative incidence function in the presence of competing risks (for example, Coviello and Boggess [2004]; Hinchliffe and Lambert [2013])
  - RS framework
    - Life-table approach (Cronin and Feuer 2000) implemented in `strs`
    - Model-based approach (Lambert et al. 2010) implemented in `stpm2cm`
- Net survival (measure)
  - Cause-specific framework
    - Censored survival times of those who die of causes other than cancer and application of standard estimators (for example, Kaplan–Meier)
  - RS framework
    - Ederer I estimator (Ederer, Axtell, and Cutler 1961)
    - Ederer II estimator (Ederer and Heise 1959)
    - Hakulinen estimator (Hakulinen 1982)
    - Model-based estimator (for example, Estève et al. [1990]; Lambert and Royston [2009])
    - Pohar Perme estimator (Pohar Perme, Stare, and Estève 2012)

For population-based cancer survival, we typically estimate net survival in an RS framework, and this is our primary focus in this article. The `strs` command supports each of the five estimators listed above for net survival in an RS framework. The Pohar Perme estimator estimates net survival in an RS framework but is not RS (see section 3.4 for details).

The choice of measure, setting, and estimator depends on the specific research question and the available data. The cause-specific framework requires accurate classification of cause of death, whereas the RS framework requires appropriate estimation of the expected survival (which usually requires an assumption that the patients would be similar to the general population if they did not have cancer). Cancer patients in clinical trials are usually specially selected and, therefore, not representative of the general population. The classification of cause of death is usually quite good in clinical trials, so cause-specific methods are usually preferred.

For population-based studies, on the other hand, cause of death comes from routine information from the death certificate (if it is available at all). For many types of cancer, it is reasonable to assume that the patients would have mortality similar to the general population if they did not have cancer, so an RS framework is usually preferred. There are scenarios, however, where it may be preferable to estimate net survival in a cause-specific setting (see Howlader et al. [2010] for examples). Similarly, there are some research questions in population-based cancer survival that are best addressed using crude survival rather than net survival (Eloranta et al. 2013).

3 Estimating net survival in an RS framework

Among the five estimators of net survival in an RS framework, the Ederer II, Pohar Perme, and model-based estimators all have practical merit.[1] Although the Ederer II method is theoretically biased, the bias in age-standardized estimates is usually negligible in practice. The Pohar Perme estimator has a slightly higher variance than the other two methods, but at five years, there is usually very little difference between the three preferred approaches (Seppä, Hakulinen, and Pokhrel Forthcoming; Lambert, Dickman, and Rutherford 2014).

3.1 RS

RS is defined as the ratio of the all-cause survival of the patients, $S_i(t)$, to the all-cause survival that would be expected, $S_i^*(t)$, in the absence of the specific disease under study. We note that RS is always estimated in an RS framework, but not all estimators in an RS framework are RS. The first three estimators in an RS framework in section 2 are RS, whereas the last two are not. It is important to recognize the distinction, made clear by Pohar Perme, Stare, and Estève (2012), between RS (the ratio of the marginal observed

1. The Ederer I and Hakulinen estimators, on the other hand, are not recommended; we include them here primarily for academic purposes.

to the marginal expected survival) and net survival (the average of the individual-specific RS). That is,

$$\text{RS} = \frac{\frac{1}{n}\sum_{i=1}^{n} S_i(t)}{\frac{1}{n}\sum_{i=1}^{n} S_i^*(t)}$$

is not necessarily equal to

$$\text{net survival} = \frac{1}{n}\sum_{i=1}^{n} \frac{S_i(t)}{S_i^*(t)}$$

Nevertheless, RS was developed as an estimator of net survival, and we maintain that it should still be considered an estimator of net survival (albeit with a small bias). As recognized long ago, RS is not a perfect estimator of net survival, hence the various approaches to estimating the marginal expected survival (described in section 3.3). The concept of RS was introduced by Berkson (1942), although he did not use the term RS. Ederer, Axtell, and Cutler (1961) defined the "RS rate" as

> the ratio of the observed survival rate in a group of patients, during a specified interval, to the expected survival rate. The expected survival rate is that of a group similar to the patient group in such characteristics as age, sex, and race, but free of the specific disease under study.

We note that RS is a ratio, not a rate, and that observed and expected survival are proportions rather than rates, but we otherwise use this same definition.

Berkson (1942) proposed RS as an estimator for "the survival so far as cancer is concerned", which is the concept today known as net survival. Although Ederer and colleagues did not use the term "net survival", it is clear they viewed both cause-specific survival and relative survival as estimators of net survival, a view that we share. Estève et al. (1990) also expressed this view and wrote "although this is hardly explicit in Ederer, Axtell, and Cutler (1961), the intention of the originators of this concept was to estimate net survival".

3.2 Estimating observed (all-cause) survival

For traditional cohort life tables, `strs` uses 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 alive at the start of interval i, d_i is the number of deaths in the interval, and $l'_i = l_i - w_i/2$ is the effective number at risk (l_i is the number alive at the start of the interval and w_i is the number censored during the interval). In period analysis (see section 4.6), survival times can be left-truncated in addition to being right-censored, so fewer subjects are at risk for the full interval. In this case, w_i would need to represent the number of individuals whose survival time was left-truncated or right-censored.

As such, whenever late entry is detected (that is, a period approach is used), `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 is 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, $p_i = \exp\{k_i \times (-d_i/y_i)\}$.

Because 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.

3.3 Estimating expected survival

The three most widely known methods for estimating expected survival for the purpose of estimating RS are the Ederer I (Ederer, Axtell, and Cutler 1961), Ederer II (Ederer and Heise 1959), and Hakulinen (Hakulinen 1982) estimators. The Hakulinen estimator was recommended until 2011, when Hakulinen suggested converting to the Ederer II estimator (Hakulinen, Seppä, and Lambert 2011). That is, the Ederer II estimator is the preferred method among the estimators of relative survival; the Ederer I and Hakulinen estimators are described primarily for academic purposes. The Ederer I estimator is, however, useful for purposes other than estimating RS.

`strs` implements all three methods, with Ederer II being the default. 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 matched individual is considered to be at risk for the purpose of estimating expected survival.[2]

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.

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 RS ratio (Hakulinen 1982) because the method does not allow for the possibility that potential follow-up

2. The mathematical details of the methods are available in an appendix to this publication at http://www.pauldickman.com/rsmodel/expected.pdf.

times of the patients may be unequal lengths. See Hakulinen, Seppä, and Lambert (2011); Pohar Perme, Stare, and Estève (2012); Rutherford, Dickman, and Lambert (2012); Danieli et al. (2012); Seppä, Hakulinen, and Pokhrel (Forthcoming); and Lambert, Dickman, and Rutherford (2014) for further details of the differences between the approaches.

3.4 Pohar Perme estimator of net survival

Pohar Perme, Stare, and Estève (2012) showed that the Ederer I, Ederer II, and Hakulinen estimators of net survival were biased, and they described a new, unbiased estimator. With the new approach, called the Pohar Perme estimator, net survival for a cohort is estimated by weighting by the inverse of the individual-specific expected survival probabilities. The weights inflate the observed person-time and number of deaths to account for person-time and deaths not observed as a result of mortality due to competing causes.

The Pohar Perme estimator (Pohar Perme, Stare, and Estève 2012) was developed for continuous survival times, yet cancer registries often have only discrete survival times (for example, survival time in completed months or completed years). Therefore, we have implemented the Pohar Perme approach in a life-table framework that is suitable when survival times are discrete but works equally well when survival times are continuous (estimates are essentially identical to Pohar Perme's R command). The `strs` command implements two alternative approaches to estimation: actuarial and hazard transformation, which give similar results.

By default, an actuarial approach is used for estimation where weights are based on the cumulative expected survival at the midpoint of the interval. If late entry is detected (that is, period analysis) or the `ht` option is specified, then net survival is estimated by transforming the cumulative excess hazard. The algorithm for the hazard transformation approach is identical to that implemented in `stnet` and described in a companion article (Coviello et al. 2015). The estimates obtained by `strs` and `stnet` are identical, but `stnet` is slightly faster because it is optimized for the one estimator.

Our actuarial estimator of net survival, NS_i, is

$$\text{NS}_i = \frac{1 - \frac{d_i^w}{n_i^w - c_i^w/2}}{\exp\left\{ -\frac{\sum_j^{n_i} \lambda_j^{*w} - \sum_j^{c_i} \lambda_j^{*w}/2 - \sum_j^{d_i} \lambda_j^{*w}/2}{n_i^w - (d_i^w + c_i^w)/2} \right\}}$$

where n_i^w, d_i^w, and c_i^w are the weighted number of individuals alive at the start of the interval, weighted number of deaths during the interval, and weighted number of censorings during the interval, respectively. λ_j^{*w} is the weighted expected hazard. The weights are the inverse of the cumulative expected survival probability and are computed at the midpoint of each interval i.

3.5 Standard errors and confidence intervals

The standard error of the observed survival proportion is estimated using Greenwood's (1926) method when life-table estimation is used. When it is estimated using a hazard transformation approach, the variance of the cumulative hazard (Breslow and Day 1987, equation 2.2) is

$$\mathrm{var}(\Lambda) = \sum_{\text{intervals } i} \frac{k_i^2 d_i}{y_i^2}$$

where k_i is the interval width, d_i is the number of deaths, and y_i is the person-time at risk. When using the delta method, the variance of the survival proportion is

$$\begin{aligned} \mathrm{var}(S) &= \mathrm{var}\{\exp(-\Lambda)\} \\ &= \left\{ \frac{d}{d\Lambda} \exp(-\Lambda) \right\}^2 \mathrm{var}(\Lambda) \\ &= S^2 \mathrm{var}(\Lambda) \end{aligned}$$

The standard error of the RS ratio is estimated as the standard error of the observed survival proportion divided by the expected survival proportion (Ederer, Axtell, and Cutler 1961).[3] Confidence intervals (CIs) are calculated on the log cumulative-hazard scale; that is, we first calculate a CI for $\log(-\log S)$, and then we back-transform to the survival scale.

The `strs` command calculates standard errors and CIs, but they are sometimes suppressed in the display of results in this article. A bias–variance tradeoff is involved in some of the choices facing practitioners of these methods. For example, period analysis (see section 4.6) excludes person-time at risk (and, hence, increases variance) to obtain up-to-date estimates. In addition to the choice between period and cohort approaches, a bias–variance tradeoff also exists in the choice of the width of the period window. Lambert, Dickman, and Rutherford (2014) discuss the bias–variance tradeoff in choosing between the various estimators of net survival in an RS framework.

4 The strs command

In general, two data files are required 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 (`popmort.dta`; see section 4.3). 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

3. This is standard, 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 RS ratio (usually overestimation due to positive correlation between the standard errors of the observed and expected survivals).

do not have to be of equal lengths). The expected probabilities are then obtained by merging with the `popmort.dta` file, and the data are collapsed to obtain one observation for each life-table interval. Expected survival may be estimated using the Ederer I (`ederer1` option), Ederer II (default), or Hakulinen method (`potfu()` option).

4.1 Syntax

`strs using` *filename* [*if*] [*in*] [*weight*], `mergeby(`*varlist*`) breaks(`*numlist*`)`
[`by(`*varlist*`) diagage(`*varname*`) diagyear(`*varname*`) attage(`*newvar*`)`
`attyear(`*newvar*`) survprob(`*varname*`) maxage(`#`) potfu(`*varname*`) ederer1`
`pohar ht calyear cuminc standstrata(`*varname*`) brenner list(`*varlist*`)`
`keep(`*varlist*`) format(%`*fmt*`) notables level(`#`) save`[`(replace)`]
`savstand(`*filename*[`, replace`]`) savind(`*filename*[`, replace`]`)`
`savgroup(`*filename*[`, replace`]`) ]`

`using` *filename* specifies a file containing general-population survival probabilities (see section 4.3).

Importance weights (`iweights`) can be used to produce age-standardized estimates; see the example in section 4.7.

4.2 Options

`mergeby(`*varlist*`)` specifies the variables that uniquely determine the records in the file of general-population survival probabilities (the `using` file, also known as the `popmort.dta` file). The `using` file must be sorted by these variables because the patient file and `using` file are merged according to these variables. `mergeby()` is required.

`breaks(`*numlist*`)` specifies the cutpoints for the life-table intervals as an ascending *numlist* commencing at 0. The cutpoints need not be integers nor equidistant, but the units must be years; for example, specify `breaks(0(0.0833)5)` for monthly intervals up to five years. `breaks()` is required.

`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, which does not have to contain integer values. The default is `diagage(age)`.

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

`attage(`*newvar*`)` specifies the variable containing attained age (that is, 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. The default is `attage(_age)`.

`attyear(`*newvar*`)` specifies the variable containing attained calendar year (that is, 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. The default is `attyear(_year)`.

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

`maxage(#)` 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. The default is `maxage(99)`.

`potfu(`*varname*`)` specifies the variable containing the last time of potential follow-up. This option is required for calculating Hakulinen estimates of expected survival and causes `strs` to report Hakulinen estimates by default. The 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 4.5.

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

`pohar` specifies that the Pohar Perme estimates of net survival be calculated and causes `strs` to report these by default (unless `potfu()` is also specified).

`ht` specifies that survival be estimated by transforming the estimated cumulative hazard. `ht` can be specified with Ederer II (the default), Hakulinen (`potfu()`), and Pohar Perme (`pohar`), but not with Ederer I (`ederer1`). The hazard transformation approach is the default when late entry is detected (for example, period analysis); otherwise, survival is estimated using an actuarial approach.

`calyear` causes `strs` to split follow-up by each calendar year, resulting in slightly more accurate estimates but at the expense of computational efficiency. `calyear` is available for use only with the `pohar` option or Ederer II estimation (the default).

`cuminc` specifies that cumulative incidence of death due to cancer (`ci_dc`) and cumulative incidence of death due to causes other than cancer (`ci_do`) be calculated using the method of Cronin and Feuer (2000). Note that the cumulative incidence of death due to cancer is estimated in the presence of competing risks, so it will be lower than (1−RS) because the latter is assumed to be in the absence of competing risks.

`standstrata(`*varname*`)` specifies a variable defining strata across which to average the cumulative survival estimate. With this option, a *weight* must also be specified as follows: `[iweight=`*varname*`]`.

`brenner` specifies that the age standardization be performed using the approach proposed by Brenner et al. (2004). This option requires that `standstrata()` (and, therefore, `[iweight=`*varname*`]`) is also specified.

`list(`*varlist*`)` specifies the variables to be listed in the life tables. The variables `start` and `end` are included by default; however, if only one of these is specified in the `list()` option, then the other is suppressed.

`keep(`*varlist*`)` restricts the variables to be written to the individual-level output dataset (named `individ.dta` by default). This option requires that `save()` or `saveind()` is also specified.

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

`notables` suppresses display of the life tables.

`level(#)` sets the confidence level based on the value of global macro `S_level`. The default is `level(95)`.

`save[ (replace) ]` creates two output datasets: `individ.dta` and `grouped.dta`. The `individ.dta` dataset contains one observation for each patient for each life-table interval, and `grouped.dta` contains one observation for each life-table interval. Use `save(replace)` to overwrite these files. Excess mortality (RS) may be modeled using these output datasets (see section 5).

`savstand(`*filename*`[ , replace ])` specifies that standardized estimates be saved to an output dataset.

`savind(`*filename*`[ , replace ])` and `savgroup(`*filename*`[ , replace ])` specify alternative filenames for the individual and grouped output datasets, respectively.

4.3 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 those variables which uniquely determine the records and upon which it is assumed that expected survival depends. Typically, those variables are age, sex, and period, but further variables may be included, such as race, region of residence, or social class (Coleman et al. 1999). Such probabilities (or corresponding rates that can be transformed to probabilities) are available from the Human Mortality Database[4] for many populations or can be obtained from local government authorities (typically the central statistics office). The filename is specified via the `using` option, and the variables by which the file is sorted are specified using the `mergeby(`*varlist*`)` option. The following is a listing of the first five rows of the Finnish `popmort.dta` file:

4. See http://www.mortality.org/.

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

	sex	_year	_age	prob	rate
1.	1	1951	0	.96429	.0363632
2.	1	1951	1	.99639	.0036165
3.	1	1951	2	.99783	.0021724
4.	1	1951	3	.99842	.0015812
5.	1	1951	4	.99882	.0011807

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—`popmort.dta` 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.dta` file, on the other hand, should contain variables `_age` and `_year` because the expected probabilities are merged using these "time-updated" variables. Alternative variable names can be specified using the `attage()` and `attyear()` options.

4.4 Example 1: Life-table estimates of RS

Here we illustrate the command using data provided by the Finnish Cancer Registry on patients diagnosed with colon carcinoma in Finland, 1975–1994. These data are distributed with the `strs` package along with do-files to reproduce all analyses presented in this article. We first estimate life tables for each gender (we show only the table for males) using patients with clinically localized (`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
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. generate id = _n
. quietly stset surv_mm, failure(status==1 2) id(id) scale(12)
. strs using popmort if stage==1, breaks(0 0.5 1(1)10) mergeby(_year sex _age)
> by(sex) list(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

(*output omitted*)

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 RS (`cr_e2`). The estimated one-year RS ratio is 0.922, and the estimated five-year relative-survival ratio is 0.770. Other quantities provided by default but omitted here (using the `list()` option) because of space limitations are interval-specific observed survival (`p`), interval-specific expected survival (`p_star`), interval-specific RS (`r`), and 95% CIs for the interval-specific RS ratio. A variable name commencing with `c` typically indicates a cumulative estimate rather than an interval-specific estimate.

When we `stset` the data, all deaths are classified as events (values `1` and `2` of the variable `status` in these data indicate death due to cancer and noncancer, 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 used 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 applied a scale factor of `scale(12)`. We could have also used variables containing dates of diagnosis (`dx`) and exit (`exit`) to `stset` the data (see the next example).

Because the life-table estimates can be saved to a dataset (see the `save()` option), it is simple to produce graphs or tables of quantities of interest. For example, the following example illustrates how we can tabulate the number of patients initially at

risk along with the five-year observed, expected, and RS for each combination of age and sex. Summary tables such as these are often presented in cancer registry reports and scientific publications.

```
. use colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. generate id = _n
. quietly stset surv_mm, failure(status==1 2) id(id) scale(12)
. strs using popmort if stage==1, breaks(0(1)10) mergeby(_year sex _age)
> by(sex agegrp) save(replace) notable
        failure _d:  status == 1 2
  analysis time _t:  surv_mm/12
                id:  id
No late entry detected - p is estimated using the actuarial method
. use grouped, clear
(Collapsed (or grouped) survival data)
. bysort sex agegr (end): generate n0=n[1]
. list sex agegr n0 cp cp_e2 cr_e2 lo_cr_e2 hi_cr_e2 if end==5, sepby(sex) noob
```

sex	agegrp	n0	cp	cp_e2	cr_e2	lo_cr_e2	hi_cr_e2
Male	0-44	161	0.7737	0.9817	0.7881	0.7102	0.8486
Male	45-59	462	0.7686	0.9335	0.8233	0.7766	0.8636
Male	60-74	1228	0.5945	0.7915	0.7512	0.7128	0.7878
Male	75+	769	0.4131	0.5312	0.7777	0.7067	0.8479
Female	0-44	136	0.7657	0.9932	0.7709	0.6866	0.8358
Female	45-59	531	0.7765	0.9763	0.7953	0.7536	0.8314
Female	60-74	1488	0.6993	0.8883	0.7873	0.7588	0.8141
Female	75+	1499	0.4854	0.6210	0.7816	0.7374	0.8249

We see that the five-year observed survival (`cp`) decreases with age (as expected) but the five-year RS (`cr_e2`) is similar across categories of age and sex. We could also use the data in `grouped.dta`, for example, to plot survival estimates as a function of follow-up time (see figure 1 in section 5.2 for a more advanced example).

4.5 Example 2: RS and net survival using four different methods

We now estimate RS using the three previously discussed methods for estimating expected survival (see section 3.3) and using the Pohar Perme estimator of net survival. To estimate expected survival using the Hakulinen method, we use the `potfu()` option to specify a variable containing the last date of potential follow-up for each patient. If the `ederer1` option is specified, then Ederer I estimates of expected survival and RS are provided. The `pohar` option instructs `strs` to calculate Pohar Perme estimates of net survival (see section 3.4). Ederer II estimates are produced by default (no option is required). The following example illustrates how all estimates can be tabulated:

```
. use colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. generate id = _n
. quietly stset exit, origin(dx) failure(status==1 2) id(id) scale(365.24)
. generate long potfu = date("31/12/1995","DMY")
. strs using popmort if stage==1, breaks(0(1)10) mergeby(_year sex _age)
> by(sex) list(n d w cr_e1 cr_e2 cr_hak cns_pp) ederer1 potfu(potfu) pohar
  (output omitted)
-> sex = Male
```

start	end	n	d	w	cr_e1	cr_e2	cr_hak	cns_pp
0	1	2620	328	0	0.9238	0.9238	0.9238	0.9212
1	2	2292	229	166	0.8758	0.8732	0.8756	0.8705
2	3	1897	180	139	0.8361	0.8312	0.8359	0.8300
3	4	1578	140	119	0.8050	0.7986	0.8049	0.7994
4	5	1319	113	104	0.7787	0.7715	0.7787	0.7771
5	6	1102	102	81	0.7486	0.7407	0.7487	0.7383
6	7	919	71	71	0.7333	0.7239	0.7335	0.7184
7	8	777	59	72	0.7200	0.7095	0.7202	0.7023
8	9	646	49	62	0.7082	0.6961	0.7082	0.6900
9	10	535	33	58	0.7085	0.6948	0.7087	0.6921

(*output omitted*)

We see only small differences between the estimates made using the Ederer I (`cr_e1`), Ederer II (`cr_e2`), Hakulinen (`cr_hak`), and Pohar Perme (`cns_pp`) methods. Differences between the methods are generally small during the first 10 years of follow-up.

4.6 Example 3: Cohort, complete, period, and hybrid estimation approaches

In this section, we demonstrate how to obtain period and hybrid estimates of RS. We estimate 10-year survival of patients diagnosed with localized (`stage==1`) colon carcinoma in Finland by using the Hakulinen method for estimating expected (and relative) survival (table 1, at the end of this section). Our dataset includes all patients diagnosed in 1975–1994, with follow-up until the end of 1995. We adopt the terminology for these approaches (cohort, complete, period, and hybrid) from Brenner et al. (2004). The fundamental difference between the various approaches is in the definition of person-time at risk. The call to `strs` is similar for each approach.

Cohort approach

To estimate 10-year survival using what Brenner et al. (2004) call the "cohort approach", all patients must have a potential follow-up of at least 10 years. Our dataset includes patients diagnosed in 1975–1994 with follow-up until the end of 1995. Therefore, only patients diagnosed in 1985 or earlier can contribute to the cohort estimate of 10-year survival. This is easily implemented in Stata.

```
. use colon, clear
. generate id = _n
. stset exit, origin(dx) failure(status==1 2) id(id) scale(365.24)
. generate long potfu = date("31/12/1995","DMY")
. strs using popmort if stage ==1 & yydx < 1986, breaks(0(1)10)
> mergeby(_year sex _age) by(sex) potfu(potfu)
```

Such estimates, based on patients diagnosed at least 10 years in the past, clearly will 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 what Brenner et al. (2004) call the "complete approach", although it is often referred to as the cohort approach. To estimate 10-year survival, we must include some patients 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 (that is, as recent as possible) but must, at a minimum, include patients diagnosed as far back as 1985. 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, breaks(0(1)10)
> mergeby(_year sex _age) by(sex) potfu(potfu)
```

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 life-table estimates of patient survival could be made using a period rather than a cohort (complete) approach (Brenner et al. 2004; 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 on the survival experience of only recently diagnosed patients. There is a trade-off between precision and recency; a narrow period window (for example, one year) will improve recency but reduce precision compared with a wider period window (for example, five years).

Period analysis provides 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 and Hakulinen 2009). Our approach to period estimation 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)) failure(status==1 2)
> id(id) scale(365.24) exit(time mdy(12,31,1994))
```

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

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

or Hakulinen estimates,

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

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

Applying 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 dataset contains patients diagnosed from 1989 to December 1994 with follow-up until December 1995. For this reason, we censored the follow-up of all individuals on 31 December 1994 in the previous example.

What would we do if we wanted to perform period analysis with a window between 1 January 1991 and 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 would contain contributions from patients diagnosed 1988–1993. All conditional estimates contain contributions from six potential years of diagnosis, apart from the first year, which only contains contributions from five 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 1 January 1990 to 31 December 1995 so that patients 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.

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

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 corresponds 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 RS estimates (Hakulinen method) for patients diagnosed with colon carcinoma according to the four different approaches.

Table 1. Ten-year RS (Hakulinen method) for patients diagnosed with localized colon carcinoma in Finland in 1985–1994 using four different approaches

Approach	RS_{males}	$\text{RS}_{\text{females}}$
Cohort	0.6848	0.7034
Complete	0.7127	0.7488
Period	0.7094	0.7880
Hybrid	0.7415	0.7840

4.7 Example 4: Age-standardized RS estimates

In this section, we will discuss age standardization, although one may standardize on factors other than age. Age standardization can be used to facilitate comparisons of RS between different populations, such as patients diagnosed in different calendar periods. Although RS estimates are automatically adjusted for differences in expected survival due to differing age distributions, they are not adjusted to account for the possibility that RS (excess mortality) depends on age.

Hakulinen (1977) suggested that one should consider using age standardization even when estimating RS for a single population where there is no interest in making comparisons (referred to as internal standardization or standardization using an internal standard). This is traditional direct standardization using an internal standard. We now know that internal standardization is crucial for minimizing bias when using, for example, the Ederer II estimator. RS is actually an unbiased estimator of net survival if all individuals have the same expected survival (see the two expressions in section 3.1). Such a scenario never occurs in practice, but we also know that the size of the bias is proportional to the degree of heterogeneity in expected survival. If we estimate RS

for patients within narrow age groups, as is done when age standardizing, then the expected survival of these patients will be similar and, therefore, the bias in RS will be much smaller than if we estimate RS for all patients combined.

Table 2 shows all-age and age-specific estimates of 10-year survival for patients diagnosed with colon carcinoma in Finland, 1985–1994.

Table 2. Age-specific numbers of patients (n_i) and estimates of 10-year RS (RS_i) for patients diagnosed with colon carcinoma in Finland, 1985–1994

age (i)	n_i	RS_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
All-age	9087	0.4358	
Age-standardized		0.4324	

If we directly age standardize 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 (see table 2, above). The age-standardized 10-year RS is given by $\sum_i \text{RS}_i w_i / \sum_i w_i = 0.4324$. Specifying the `standstrata()` option causes `strs` to first produce stratified life tables for each level of the variables specified in `standstrata()` and then produce standardized estimates using the weights contained in the variable specified in `iweights`.

```
. use colon, clear
. generate id = _n
. stset exit, origin(dx) failure(status==1 2) id(id) scale(365.24)
. tab agegrp if yydx>1984
. recode agegrp 0=0.041928 1=0.147353 2=0.407065 3=0.403654, generate(standw)
. strs using popmort [iw=standw] if yydx > 1984, breaks(0(1)20)
> mergeby(_year sex _age) standstrata(agegrp) notables
```

The weights should be specified as proportions. In this example, the crude and internally age-standardized estimates were similar, although this is not always the case (Hakulinen 1977). It is possible to use the `by()` option with `standstrata()` to produce, for example, age-standardized estimates for each calendar period. For example, the following code produces age-standardized estimates for each period, using the age structure for the latter period as the standard. The variable `year8594` is an indicator 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, generate(standw)
. strs using popmort [iw=standw], breaks(0(1)20) mergeby(_year sex _age)
> standstrata(agegrp) by(year8594) notables
```

Rather than weighting based on the age distribution at the start, Brenner and Hakulinen (2004) suggest using weights that change throughout follow-up time. This is achieved by assigning individual weights to each patient and constructing a weighted life table (Brenner and Hakulinen 2004). Specifying the `brenner` option (together with `iweight`) causes `strs` to produce standardized estimates using this alternative method. With this method, if we use the actual age distribution of the patients as the standard population, then the age-standardized estimates will, unlike with the traditional method, be identical to the crude estimates (see table 3).

Table 3. Crude, age-standardized, and age-adjusted (alternative) 10-year RS estimates obtained in each period for patients with colon carcinoma in Finland, 1975–1994; the age distribution for 1985–1994 is used as the standard population

	10-year RS		
Period	Crude	Age-standardized (traditional)	Age-adjusted (alternative)
1975–1984	0.4035	0.4023	0.3998
1985–1994	0.4358	0.4324	0.4358

The two groups under comparison have a very similar age structure, so there are only small differences between the different approaches, but this is not always the case (Brenner and Hakulinen 2004). The same technique can be used with respect to other factors, such as race or stage, but modeling is generally the method of choice for comparing survival between populations after adjustment for multiple covariates. See Pokhrel and Hakulinen (2008, 2009) for an overview of the various approaches for age standardizing RS and how they should be interpreted.

4.8 Example 5: Estimating crude probabilities of death

Both RS and cause-specific survival estimate the same underlying hypothetical quantity: net survival—the probability of survival where the specific cancer is the only possible cause of death. Cause-specific survival estimates it directly whereas RS estimates it by estimating excess mortality. A 15-year RS of 60%, for example, implies patients have a 60% probability of surviving 15 years or more following diagnosis in the hypothetical scenario where the cancer of interest is the only possible cause of death. The net probability of death due to cancer within 15 years is $100 - 60 = 40\%$ and is interpreted under the assumption that patients cannot die of other causes.

Net survival is extremely useful for etiological or public health research, where we may wish, for example, to compare survival over time or between groups of patients while correcting for differences in noncancer mortality. Patients, however, do not live in this hypothetical world and estimates of crude mortality or crude survival may be of greater

interest.[5] That is, the crude probability of dying of cancer within 15 years is the actual probability of dying in the presence of competing risks and will be lower than the net probability of dying of cancer. Cronin and Feuer (2000) showed how crude probabilities of death due to cancer and due to causes other than cancer can be estimated from life tables. Their approach is implemented by `strs` using the `cuminc` option. Lambert et al. (2010) showed how these quantities can be estimated using models for excess mortality and then implemented the approach with `stpm2cm`.

```
. strs using popmort if age>74, breaks(0(1)10)
> cuminc mergeby(_year sex _age) list(cr_e2 ci_dc ci_do)
        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
```

start	end	cr_e2	ci_dc	ci_do
0	1	0.5994	0.3816	0.0760
1	2	0.5104	0.4584	0.1228
2	3	0.4772	0.4842	0.1626
3	4	0.4527	0.5015	0.1991
4	5	0.4413	0.5086	0.2329
5	6	0.4253	0.5173	0.2640
6	7	0.4162	0.5218	0.2923
7	8	0.4092	0.5246	0.3183
8	9	0.3997	0.5280	0.3418
9	10	0.4004	0.5278	0.3632

In the output above, 1 minus `cr_e2` is the net probability of death due to cancer, while `ci_dc` and `ci_do` are the crude probabilities of death (also known as cumulative incidence) due to cancer and due to other causes, respectively. That is, during a 10-year follow-up of these patients who were aged 75 or over at diagnosis, we estimate that 53% will have died of cancer, 36% will have died of causes other than cancer, and 11% will be alive. In the hypothetical scenario where patients can die only of cancer, we estimate that 60% of patients will have died of cancer and 40% of patients will not have died of their cancer within 10 years.

5 Modeling excess mortality

The `strs` command was specifically designed to facilitate modeling. It produces two output datasets that can be used for modeling. The mortality analogue of RS is excess mortality, and it is this quantity that is modeled. The total hazard at the time since diagnosis, t, for persons diagnosed with cancer (with covariate vector $\mathbf{z}$) is modeled as

5. Note that the term "crude survival" is often used to mean "all-cause survival", although here we use the term "crude" as it is used within the theory of competing risks (Tsiatis 2005), and we use "observed survival" as a synonym for all-cause survival.

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,

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

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 $\mathbf{z}$, 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 tumor-specific covariates, such as histology or stage. For simplicity, we write that the expected hazard is a function of $\mathbf{z}$, even though it does not vary over all elements of $\mathbf{z}$.

We partition the follow-up time into bands corresponding to life-table intervals. These are typically one year in length, although it is possible to use shorter intervals early in the follow-up, where mortality is often higher and changing rapidly (as in section 4.4). We also construct a set of indicator variables (one indicator variable for each interval, excluding the reference interval) and incorporate it into the covariate matrix. We 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 RS model is, therefore, written as

$$\lambda(\mathbf{x}) = \lambda^*(\mathbf{x}) + \exp(\mathbf{x}\beta) \tag{1}$$

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 (1) is the assumption that the excess hazards for any two patient subgroups are proportional over follow-up time. We can, however, incorporate nonproportional excess hazards by including time-by-covariate interaction terms in the model. The exponentiated parameter estimates can be interpreted as excess hazard ratios, sometimes known as relative excess risks (Suissa 1999). An excess hazard ratio of, for example, 1.5 for males compared with females implies that the excess mortality associated with a diagnosis of cancer is 50% higher for males than females.

5.1 Modeling excess mortality using a full-likelihood approach

Estève et al. (1990) described a method for fitting the model in (1) directly from individual-level data using a maximum likelihood approach. The likelihood function is

$$L = \prod_{i=1}^{n} \exp\left\{-\int_0^{t_i} \lambda(s)\,ds\right\} \{\lambda(t_i)\}^{d_i}$$

where t_i is the survival time and d_i is 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

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

Although the model is specified in continuous time, it is assumed (as with all approaches described here) that the hazard is constant within prespecified bands of time, and the excess hazard $\nu(t)$ is written as $\exp(\mathbf{x}\beta)$. Fitting the model is simplified if each observation is split into separate observations for each band of follow-up. The contribution of the ijth subject band to the total log likelihood is

$$l_{ij}(\beta) = [d_{ij} \ln\{\lambda^*(\mathbf{x}_{ij}) + \exp(\mathbf{x}_{ij}\beta)\} - y_{ij} \exp(\mathbf{x}_{ij}\beta)] \tag{2}$$

where y_{ij} is the time spent by subject i in time-band j.

The Stata `ml` command with the `lf` method can be used to maximize the log-likelihood function shown in (2). The likelihood used by the `ml` command is defined in `esteve.ado`, which is part of the `strs` package and reproduced below.

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

The global macro `ML_y1` contains d, the death indicator.

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` with `by(sex year8594 agegrp)`. This causes these variables to be included in the output file (`individ.dta`), which will contain one observation for each individual for each life-table interval, and it also generates the grouped data (`grouped.dta`) by all combinations of `sex`, `year8594`, and `agegrp`. We require λ^* (the expected mortality rate) but our `popmort.dta` file contains p^*, the probability of surviving, so we transform the probability to a rate (see section 4.3).

```
. use colon, clear
(Colon carcinoma, all stages, Finland 1975-94, follow-up to 1995)
. generate id = _n
. quietly stset surv_mm, failure(status==1 2) id(id) scale(12)
. strs using popmort if stage==1, breaks(0(1)10) mergeby(_year sex _age)
> by(sex year8594 agegrp) save(replace) notable noshow
No late entry detected - p is estimated using the actuarial method
```

```
. use individ if end<6, clear
(Survival data containing individual subject-band observations)
. generate rate=-ln(p_star)
. ml model lf esteve (d y rate = i.end sex year8594 i.agegrp)
. ml maximize, eform("EHR") nolog
                                                Number of obs   =      23579
                                                Wald chi2(9)    =      72.73
Log likelihood = -5969.5775                     Prob > chi2     =     0.0000
```

	EHR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
end						
2	.8286045	.0779917	-2.00	0.046	.689015	.9964739
3	.6765733	.0727639	-3.63	0.000	.5479868	.835333
4	.5383155	.069149	-4.82	0.000	.4185008	.6924325
5	.4606403	.0690407	-5.17	0.000	.343387	.617931
sex	.9545966	.0737863	-0.60	0.548	.8203999	1.110744
year8594	.734979	.055002	-4.11	0.000	.6347102	.8510879
agegrp						
1	.8663227	.135108	-0.92	0.358	.6381604	1.17606
2	1.055003	.1508525	0.37	0.708	.7971545	1.396256
3	1.341785	.2022822	1.95	0.051	.9985251	1.803045
_cons	.0844594	.015493	-13.47	0.000	.0589531	.1210012

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 in 1985–1994 and 0 for patients diagnosed in 1975–1984. We see that patients diagnosed in the recent period are estimated to experience 27% lower excess mortality compared with 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 in excess mortality between males and females.

5.2 Modeling excess mortality using Poisson regression

The RS model (1) assumes piecewise constant hazards, which implies a Poisson process for the number of deaths in each interval. This implies that the RS model can be fit in the framework of generalized linear models using a Poisson assumption for the observed number of deaths. We assume that the number of deaths for observation j, d_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 (1) is then written as

$$\frac{\mu_j}{y_j} = \frac{d_j^*}{y_j} + \exp(\mathbf{x}\beta)$$

which can be written as

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

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 generalized 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 it is defined in `rs.ado`, which is included in the `strs` package and can be viewed by typing `viewsource rs.ado`.

Example: Poisson regression

The example in section 5.1 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. Here we fit the Poisson regression model to the grouped data; if we fit the model to the data in `individ.dta`, we would obtain identical estimates to the full-likelihood approach (section 5.1) because we would be maximizing the same likelihood using the same data.

```
. use grouped if end<6, clear
(Collapsed (or grouped) survival data)
. glm d i.end i.sex i.year8594 i.agegrp, fam(pois)
> link(rs d_star) lnoffset(y) eform nolog
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	exp(b)	OIM Std. Err.	z	P>|z|	[95% Conf.	Interval]
end						
2	.7984084	.0730515	-2.46	0.014	.6673339	.955228
3	.6230213	.0671961	-4.39	0.000	.5043086	.7696785
4	.4969433	.0645561	-5.38	0.000	.3852391	.6410374
5	.4334347	.065147	-5.56	0.000	.322838	.5819191
2.sex	.9564493	.0729823	-0.58	0.560	.8235891	1.110742
1.year8594	.7308044	.0539291	-4.25	0.000	.6323935	.8445296
agegrp						
1	.8642841	.1353083	-0.93	0.352	.635911	1.174672
2	1.071568	.1534869	0.48	0.629	.8092774	1.418869
3	1.436319	.2146593	2.42	0.015	1.071613	1.925147
_cons	.0838687	.0124017	-16.76	0.000	.0627671	.1120644
ln(y)	1	(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 fitting the model in the framework of generalized linear models is that we have access to a rich theoretical

framework and can use, for example, regression diagnostics. An advantage of fitting the model to collapsed data is that we can assess goodness of fit by using the deviance or Pearson's chi-squared statistic (provided the data are nonsparse). We see that there is evidence of lack of fit (deviance is 131.4 with 70 degrees of freedom), 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 previously assumed the hazard to be piecewise constant (that is, a step function) over follow-up time, an assumption that is not attractive from a clinical or biological perspective. We might alternatively specify narrower time bands (for example, monthly) and model the effect of follow-up using a restricted cubic spline.

```
. use colon
. generate id = _n
. stset exit, origin(dx) failure(status==1 2) id(id) scale(365.24)
. generate long potfu = date("31/12/1995","DMY")
. strs using popmort if stage==1, breaks(0(0.083333333)5) mergeby(_year sex _age)
> by(sex year8594 agegrp) potfu(potfu) save(replace) notable
. use grouped, clear
. mkspline endb = end, cubic nknots(5)
. glm d endb? i.sex i.year8594 i.agegrp, failure(poisson) link(rs d_star)
> lnoffset(y)
```

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) 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 RS for each age group for males with localized colon cancer diagnosed in 1985–1994 as well as corresponding life-table estimates (Hakulinen approach) with 95% CIs.

```
. predict xb, xb nooffset                 // excess risk
. generate r_hat = exp(-exp(xb)*0.083333) // interval-specific relative survival
. bysort sex year8594 agegrp (end) :
> generate 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)) ytitle("Relative survival")
> xtitle("Years from diagnosis") xlabel(0(1)5) ylabel(0.6(.1)1, format(%3.2f))
```

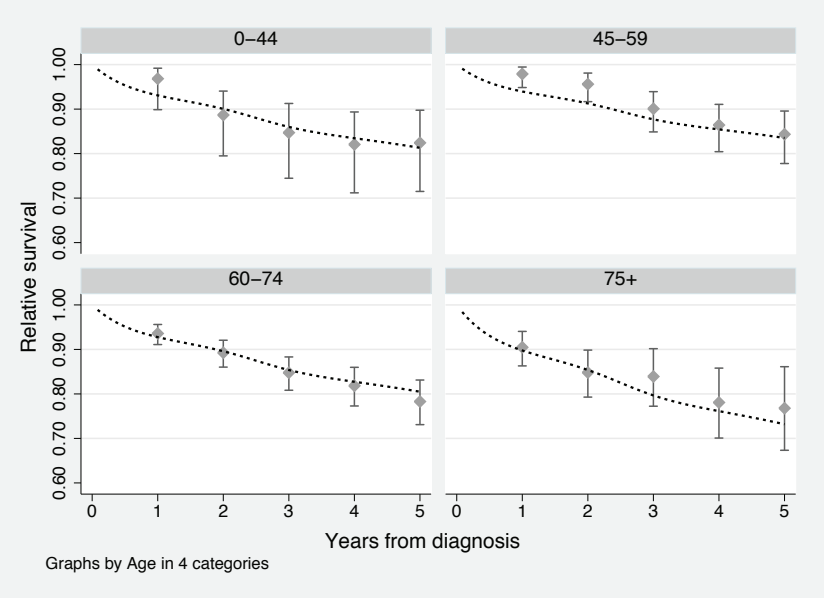

Figure 1. Model-based (dotted line) and empirical (with 95% CI) estimates of RS by age groups for males with localized colon cancer diagnosed in 1985–1994

As is common with cancer survival data, patients aged 75 years or more at diagnosis have considerably higher mortality during the first year following diagnosis, but once they have survived the first year, they experience excess mortality more similar to the other age groups. That is, the excess hazards are nonproportional by age at diagnosis.

5.3 Hakulinen–Tenkanen approach to modeling excess mortality

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

$$\ln\left(-\ln\frac{p_j}{p_j^*}\right) = \mathbf{x}\beta$$

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 is fit using the `glm` command in the usual manner.

```
. use grouped
. glm ns i.end i.sex i.year8594 i.agegrp, family(bin n_prime) link(ht p_star)
```

6 Conclusion

The `strs` command implements procedures that are commonly used in population-based cancer epidemiology. It is designed to facilitate modeling, yet its flexibility makes it useful for purposes other than modeling as well. We used it, for example, to estimate standardized incidence ratios and to estimate life expectancy. The only original theory we presented here is the actuarial estimator of net survival. The theory underlying all other approaches has been presented elsewhere, and we provided appropriate references throughout this article.

7 Acknowledgments

We gratefully acknowledge the significant contribution of Michael Hills, who coauthored the first version of this article, submitted in 2007. In particular, Michael developed the approach described in section 5.1 for modeling excess mortality by writing the log likelihood in terms of subject bands. 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. We thank Karri Seppä and Arun Pokhrel for their contribution to developing the Pohar Perme estimator in a life-table framework. The `strs` command was first released in 2004, and we thank the many users who have since then contributed suggestions and helped us correct errors.

8 References

Berkson, J. 1942. The calculation of survival rates. In *Carcinoma and Other Malignant Lesions of the Stomach*, ed. W. Walters, H. K. Gray, and J. T. Priestly, 467–484. Philadelphia: Sanders.

Brenner, H., V. Arndt, O. Gefeller, and T. Hakulinen. 2004. An alternative approach to age adjustment of cancer survival rates. *European Journal of Cancer* 40: 2317–2322.

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

Brenner, H., and T. Hakulinen. 2004. Are patients diagnosed with breast cancer before age 50 years ever cured? *Journal of Clinical Oncology* 22: 432–438.

———. 2005. Age adjustment of cancer survival rates: Methods, point estimates and standard errors. *British Journal of Cancer* 93: 372–375.

———. 2009. Up-to-date cancer survival: Period analysis and beyond. *International Journal of Cancer* 124: 1384–1390.

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: 2494–2501.

Breslow, N. E., and N. E. Day. 1987. *Statistical Methods in Cancer Research: Vol. 2—The Design and Analysis of Cohort Studies*. Lyon: IARC.

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*. London: Stationery Office.

Coviello, E., P. W. Dickman, K. Seppä, and A. Pokhrel. 2015. Estimating net survival using a life-table approach. *Stata Journal* 15: 173–185.

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

Cronin, K. A., and E. J. Feuer. 2000. Cumulative cause-specific mortality for cancer patients in the presence of other causes: A crude analogue of relative survival. *Statistics in Medicine* 19: 1729–1740.

Danieli, C., L. Remontet, N. Bossard, L. Roche, and A. Belot. 2012. Estimating net survival: The importance of allowing for informative censoring. *Statistics in Medicine* 31: 775–786.

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

Dickman, P. W., A. Sloggett, M. Hills, and T. Hakulinen. 2004. Regression models for relative survival. *Statistics in Medicine* 23: 51–64.

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 10. End Results Evaluation Section, National Cancer Institute.

Eloranta, S., J. Adolfsson, P. C. Lambert, P. Stattin, O. Akre, T. M.-L. Andersson, and P. W. Dickman. 2013. How can we make cancer survival statistics more useful for patients and clinicians: An illustration using localized prostate cancer in Sweden. *Cancer Causes Control* 24: 505–515.

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: 339–352.

Giorgi, R., M. Abrahamowicz, C. Quantin, P. Bolard, J. Estève, J. Gouvernet, and J. Faivre. 2003. A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine* 22: 2767–2784.

Greenwood, M. 1926. The errors of sampling of the survivorship table. In *Reports on Public Health and Medical Subjects, Volume 33*. 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., K. Seppä, and P. C. Lambert. 2011. Choosing the relative survival method for cancer survival estimation. *European Journal of Cancer*47: 2202–2210.

Hakulinen, T., and L. Tenkanen. 1987. Regression analyses of relative survival rates. *Applied Statistics* 36: 309–317.

Hinchliffe, S. R., and P. C. Lambert. 2013. Extending the flexible parametric survival model for competing risks. *Stata Journal* 13: 344–355.

Howlader, N., L. A. G. Ries, A. B. Mariotto, M. E. Reichman, J. Ruhl, and K. A. Cronin. 2010. Improved estimates of cancer-specific survival rates from population-based data. *Journal of the National Cancer Institute* 102: 1584–1598.

Lambert, P. C., P. W. Dickman, and M. J. Rutherford. 2014. Comparison of approaches to estimating age-standardized net survival. Submitted.

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

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

Lambert, P. C., 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: 3871–3885.

Nelson, C. P., P. C. Lambert, I. B. Squire, and D. R. Jones. 2008. Relative survival: What can cardiovascular disease learn from cancer? *European Heart Journal* 29: 941–947.

Pohar Perme, M., J. Stare, and J. Estève. 2012. On estimation in relative survival. *Biometrics* 68: 113–120.

Pokhrel, A., and T. Hakulinen. 2008. How to interpret the relative survival ratios of cancer patients. *European Journal of Cancer* 44: 2661–2667.

———. 2009. Age-standardisation of relative survival ratios of cancer patients in a comparison between countries, genders and time periods. *European Journal of Cancer* 45: 642–647.

Rutherford, M. J., P. W. Dickman, and P. C. Lambert. 2012. Comparison of methods for calculating relative survival in population-based studies. *Cancer Epidemiology* 36: 16–21.

Seppä, K., T. Hakulinen, and A. Pokhrel. Forthcoming. Choosing the net survival method for cancer survival estimation. *European Journal of Cancer*.

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

Tsiatis, A. A. 2005. Competing Risks. In *Encyclopedia of Biostatistics*, ed. P. Armitage and T. Colton, 2nd ed., 1025–1035. Chichester, UK: Wiley.

About the authors

Paul Dickman is a professor of biostatistics at Karolinska Institutet in Stockholm, Sweden. He conducts research in population-based epidemiology with a particular focus on cancer epidemiology. His main interest is in the development and application of statistical methods for studying the survival of cancer patients.

Enzo Coviello is an epidemiologist in the Unit of Statistics and Epidemiology at ASL BT in Barletta, Italy. He is a longtime Stata user and enthusiast as well as the author of some popular Stata commands, including `stcascoh`, `stcompet`, and `distrate`. His main interest is in the analysis of population-based cancer registries data.