Journal of Statistical Software

November 2018, Volume 87, Issue 8. *doi: 10.18637/jss.v087.i08*

Nonparametric Relative Survival Analysis with the R Package relsurv

Maja Pohar Perme
University of Ljubljana

Klemen Pavlič
University of Ljubljana

Abstract

Relative survival methods are crucial with data in which the cause of death information is either not given or inaccurate, but cause-specific information is nevertheless required. This methodology is standard in cancer registry data analysis and can also be found in other areas. The idea of relative survival is to join the observed data with the general mortality population data and thus extract the information on the disease-specific hazard. While this idea is clear and easy to understand, the practical implementation of the estimators is rather complex since the population hazard for each individual depends on demographic variables and changes in time.

A considerable advance in the methodology of this field has been observed in the past decade and while some methods represent only a modification of existing estimators, others require newly programmed functions. The package **relsurv** covers all the steps of the analysis, from importing the general population tables to estimating and plotting the results. The syntax mimics closely that of the classical survival packages like **survival** and **cmprsk**, thus enabling the users to directly use its functions without any further familiarization.

In this paper we focus on the nonparametric relative survival analysis, and in particular, on the two key estimators for net survival and crude probability of death. Both estimators were first presented in our package and are still missing in many other software packages, a fact which greatly hampers their frequency of use.

The paper offers guidelines for the actual use of the software by means of a detailed nonparametric analysis of the data describing the survival of patients with colon cancer. The data have been provided by the Cancer Registry of Slovenia.

Keywords: relative survival analysis, net survival, crude probability of death, R.

1. Introduction

The cause of death information in observational survival studies with long follow-up times is often incomplete or unavailable even though disease-specific information is of interest.

A typical example of such data comes from cancer registries, where only follow-up times and vital status at the end of follow-up are recorded, while cause of death is unknown or inaccurately recorded. The methodology dealing with these data has been developed under the name *relative survival analysis* – the data of the cohort are joined with the data on general population mortality that are collected by the national statistical offices. Under the assumption that the population mortality hazard is the hazard that our patients would be exposed to if they did not have the disease in question, this mortality can be used to extract the excess or cause-specific information of interest.

The idea of relative survival analysis has been introduced many years ago (Ederer, Axtell, and Cutler 1961) and has been in standard use in cancer registry data analyses. For many years, the gold standard for nonparametric estimation of survival curves has been the Hakulinen estimator (Hakulinen and Tenkanen 1987), but it has been recently shown that this estimator does not have the expected properties. This gave rise to methodological advances, either in terms of corrections of this estimator (Pokhrel and Hakulinen 2009; Hakulinen, Seppä, and Lambert 2011) or in the search for alternative measures (Cronin and Feuer 2000; Lambert, Dickman, Nelson, and Royston 2010). Many controversies in the field were resolved by the recent paper of Pohar Perme, Stare, and Estève (2012) that defined the often confused theoretical measures of interest and proposed a consistent nonparametric estimator of net survival. An overview of the different measures is given in Pohar Perme, Estève, and Rachet (2016), assumptions of the net survival measure were thoroughly studied and discussed in Pavlič and Pohar Perme (2018).

This paper is a practical complement to the recent methodological advances as it describes the functions for estimating the measures of interest in R (R Core Team 2018). In particular, it focuses on nonparametric estimation of three measures: net survival, crude probability of death and relative survival ratio. It discusses the practical problems in the implementation and usage of the estimators. The paper represents a companion to the R package – it explains the basic concepts in a currently rather confused field, states the formulae for the implemented estimators, explains the R syntax and works through an example.

All the new functions have been added to the package **relsurv** (Pohar Perme 2018; Pohar and Stare 2006, 2007) that has previously focused on regression modeling in relative survival setting. Package **relsurv** is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=relsurv`. Other R packages such as **mexhaz** (Charvat and Belot 2018) and **rstpm2** (Clements and Liu 2018) include more elaborate regression modeling options (flexible parametric parametric models, random effects, penalization). Functions for estimation in the relative survival setting are being developed in other statistical environments as well, with the work in Stata (StataCorp 2015) currently being the closest to the extent covered in R. We believe it is crucial that all the different concepts are covered in one statistical software package, since the different measures have a different interpretation and one may wish to use several or all to tell the complete story.

We limit ourselves to methods for continuous-time data and introduce formulae for the continuous-time version of the nonparametric estimator of crude probability of death, which was up to now available only for discretely reported data (interval data) (Cronin and Feuer 2000). For completeness we add also the formulae for the net estimator that were first introduced in Pohar Perme *et al.* (2012). A clear distinction should also be made between parametric and nonparametric methods; we focus on nonparametric methods in this work. Since well-defined unbiased estimators now exist for each of the measures, we have also avoided

any ad-hoc developed estimators that have no clearly defined population value regardless of whether they were shown to have reasonable properties in practice.

The paper is organized as follows. Section 2 focuses on a clear theoretical presentation of the different concepts and estimators. Section 3 presents the R functions for the described estimators and discusses some practical problems encountered when working with these estimators. Section 4 describes the usage of these functions and Section 5 describes a detailed example of the analysis with all the intermediate steps. Section 6 concludes the paper.

2. Computational methods and theory

In this work, we shall focus on three different measures, each of them carrying some information on the effectiveness of disease treatment: the relative survival ratio, the net survival and the crude probability of death.

Let $S_O(t)$ denote the overall survival, i.e., the probability that an individual is still alive. This survival is referred to as "overall" since it is calculated without respect to the cause of death – we are simply interested in the proportion of individuals still alive in the population at a certain time point. The other quantity of central importance in the relative survival field is the "expected" or the "population" survival $S_P(t)$ which is the survival curve of a group of people that matches our sample of patients in terms of the demographic variables at the time of diagnosis, but does not have the disease of interest. We assume that the value of $S_P(t)$ can be read from the population mortality tables, for N patients, the population survival equals $S_P(t) = \frac{1}{N}\sum_i S_{Pi}(t)$. In this, we assume that the deaths due to the disease in question form only a negligible part of the population mortality and that the national mortality tables would not change much if the patients having this disease were excluded from the calculation. Note here, that the naming of the measures is slightly confusing, we shall speak of the "population survival" and refer to the survival of the general population, but also speak of the measures defined on the "population" (the theoretical values) and then later discuss their estimators that are of course calculated on a sample.

The most simple measure that has been in use for years is the relative survival ratio (Ederer *et al.* 1961)

$$S_R(t) = \frac{S_O(t)}{S_P(t)}.$$

The ratio describes how our patients' survival compares to that of the general population. It is typically below 1 indicating that the survival of the patients is worse. There is no reason why this curve could not also increase, it is not a survival function of any group of patients and thus not necessarily a monotonically decreasing function (Pohar Perme *et al.* 2012). When comparing this measure between two cohorts with different demographic values, one should always take into account its relativity – even if two compared cohorts have the same disease-specific hazards, their ratio can be different, usually it is the cohort with the better population survival that has a lower relative survival ratio.

In order to define the other two measures, we assume that the overall hazard of each individual λ_{Oi} can be written as a sum of "disease-specific" or "excess" hazard λ_{Ei} and the "population" hazard λ_{Pi}, i.e., $\lambda_{Oi}(t) = \lambda_{Ei}(t) + \lambda_{Pi}(t)$. With the disease-specific hazard being of primary interest,

we wish to report a summary of λ_{Ei} through time and over individuals. To this end, we define the individual relative survival ratio as

$$S_{Ei}(t) = \exp\{-\int_0^t \lambda_{Ei}(u)du\} = \frac{\exp\{-\int_0^t \lambda_{Oi}(u)du\}}{\exp\{-\int_0^t \lambda_{Pi}(u)du\}}, \tag{1}$$

the marginal relative survival ratio of a cohort of size N is thus

$$S_E(t) = \frac{1}{N}\sum_{i=1}^{N} S_{Ei}(t).$$

Note that despite the notation (S_E), the marginal relative survival ratio is not necessarily a survival function.
Similarly to the relative survival ratio which is the ratio of averages, the net survival can be written as the average of ratios:

$$S_R(t) = \frac{\frac{1}{N}\sum_{i=1}^{N} S_{Oi}(t)}{\frac{1}{N}\sum_{i=1}^{N} S_{Pi}(t)}; \qquad S_E(t) = \frac{1}{N}\sum_{i=1}^{N} \frac{S_{Oi}(t)}{S_{Pi}(t)}. \tag{2}$$

However, contrary to the relative survival ratio, this measure is much more suitable for comparisons between cohorts with different population survival, since it is by definition not affected by the population mortality hazard (1). If two cohorts have equal disease-specific hazards, their net survival curves shall be equal.

As an alternative to the "average of ratios" interpretation, one can refer to the measure as the probability that a patient is still alive in the hypothetical world where the disease of interest is the only possible cause of death. To make it estimable from real life data, we add the assumption that the hazard λ_{Ei} remains unchanged when the other causes are removed. When using this interpretation, we refer to the measure as net survival. Such a hypothetical world is of course unreasonable and the estimation of the survival in it requires some strong untestable assumptions. The reason why we nevertheless wish to estimate this measure comes from the wish to get a measure that does not depend on the probability of dying due to other causes. This measure is therefore of use when interested in comparisons between populations with different mortality (different countries, same country in different time periods).

Net survival (or marginal relative survival ratio) is calculated whenever the disease-specific hazard is the sole quantity of interest, but we wish to express it on a survival scale.

As the third option, we consider splitting the overall mortality ($1 - S_O(t)$) into the two cumulative incidence functions: the crude probability of death from the disease in question by time t (also referred to as crude cancer mortality)

$$F_C(t) = \mathsf{P}(T \leq t, \text{death due to disease}) = \int_0^t S_O(u-)d\Lambda_C(u),$$

and the crude probability of death from other causes

$$F_P(t) = \mathsf{P}(T \leq t, \text{death due to other causes}) = \int_0^t S_O(u-)d\Lambda_P(u).$$

Here, T is the random variable denoting the time from diagnosis to the event, while Λ_C and Λ_P are the cumulative versions of the cause-specific hazards which satisfy the equation $\lambda_O(t) = \lambda_C(t) + \lambda_P(t)$ on a group level, i.e., $\lambda_C(t) = \frac{\sum_i S_{Oi}(t)\lambda_{Ei}(t)}{\sum_i S_{Oi}(t)}$ and $\lambda_P(t) = \frac{\sum_i S_{Oi}(t)\lambda_{Pi}(t)}{\sum_i S_{Oi}(t)}$ (see Pohar Perme *et al.* 2012, where notation λ_E^* and λ_P^* was used). Crude probability of death is a measure that is clearly defined in the real world, but again depends on the population mortality differences. If two cohorts have the same disease-specific hazards, the crude probability of death of the cohort with the lower population hazards may be higher: Some patients may die of other reasons before they could die from cancer in a cohort with high population hazard, whereas they would die of cancer if the population hazard was lower.

We now introduce some further notation needed to define the estimators of the above mentioned measures. Let $dN_i(t)$ count the number of events of individual i ($i = 1, \ldots, n$) at time t and $dN(t) = \sum dN_i(t)$ be the total number of events at time t. $N_i(t) = \int_0^t dN_i(s)$ is a counting process that starts at 0 and jumps to 1 at the time when the individual i dies. The at risk process is denoted by Y, we use $Y_i(t)$ as the indicator whether a person is still at risk and $Y(t) = \sum Y_i(t)$ as the total number at risk at time t. Both processes (N and Y) are observed on the cohort. The information we need from the population mortality tables is given by $\lambda_{Pi}(t)$ – for each individual, we have the population mortality hazard that they are exposed to at a certain time point. We use it to calculate the cumulative hazard

$$\Lambda_{Pi}(t) = \int_0^t \lambda_{Pi}(u)du \tag{3}$$

and the population survival function for each individual $S_{Pi}(t) = \exp\{-\Lambda_{Pi}(t)\}$.

Using the above defined quantities, the relative survival ratio estimator equals

$$\hat{S}_R(t) = \frac{\hat{S}_O(t)}{\hat{S}_P(t)}, \tag{4}$$

where $\hat{S}_O(t)$ is the estimator of the overall survival, i.e., its cumulative hazard function is estimated as

$$\hat{\Lambda}_O(t) = \int_0^t \frac{dN(s)}{Y(s)}, \tag{5}$$

and $\hat{S}_P(t) = \frac{1}{n}\sum_{i=1}^{n} S_{Pi}(t)$.

The standard error of the population mortality data is assumed negligible compared to that of the observed data, therefore, only the observed part is important for the variance estimation, i.e.,

$$\widehat{\mathsf{VAR}}(\hat{S}_R(t)) = \frac{1}{\hat{S}_P^2(t)}\widehat{\mathsf{VAR}}(\hat{S}_O(t))$$

is used to this end. For the estimator of net survival (Pohar Perme *et al.* 2012), referred to as the PP estimator later in the text, the estimator of the cumulative hazard equals

$$\hat{\Lambda}_E(t) = \int_0^t \frac{\sum_{i=1}^{n} \frac{dN_i(u)}{S_{Pi}(u)}}{\sum_{i=1}^{n} \frac{Y_i(u)}{S_{Pi}(u)}} - \int_0^t \frac{\sum_{i=1}^{n} \frac{Y_i(u)}{S_{Pi}(u)} d\Lambda_{Pi}(u)}{\sum_{i=1}^{n} \frac{Y_i(u)}{S_{Pi}(u)}}. \tag{6}$$

Its variance estimator equals

$$\widehat{\mathsf{VAR}}(\hat{\Lambda}_E(t)) = \int_0^t \frac{J(u)}{\left(\sum_{i=1}^n \frac{Y_i(u)}{S_{Pi}(u)}\right)^2} \sum_{i=1}^n \frac{dN_i(u)}{S_{Pi}^2(u)},$$

where $J(t) = I(Y(t) > 0)$ is an indicator that prevents from dividing by 0, $J(t)/Y(t)$ equals 0 if $Y(t) = 0$.

The continuous-time estimator for the crude probability of death equals

$$\hat{F}_C(t) = \int_0^t \hat{S}_O(u-) d\hat{\Lambda}_C(u), \tag{7}$$

where $d\hat{\Lambda}_C(u)$ is the estimated increase of the cause specific cumulative hazard (in small intervals, see Section 3.4 for details), calculated as the difference between $d\hat{\Lambda}_O(u)$ and $d\hat{\Lambda}_P(u)$, i.e., $d\hat{\Lambda}_C(u) = d\hat{\Lambda}_O(u) - d\hat{\Lambda}_P(u)$ with $d\hat{\Lambda}_O(u) = \frac{dN(u)}{Y(u)}$ and $d\hat{\Lambda}_P(u) = \frac{1}{Y(u)} \sum_{i=1}^n Y_i(u) d\Lambda_{Pi}(u)$:

$$d\hat{\Lambda}_C(u) = \frac{dN(u)}{Y(u)} - \frac{\sum_{i=1}^n Y_i(u) d\Lambda_{Pi}(u)}{Y(u)}.$$

In order to obtain an estimator for the variance of $\hat{F}_C(t)$, we have to define an estimator of transition probability $\mathsf{P}(T \leq t, \text{death due to disease } | T > s)$:

$$\hat{F}_C(s,t) = \int_s^t \frac{\hat{S}_O(u-)}{\hat{S}_O(s)} d\hat{\Lambda}_C(u).$$

Note that the estimator of the crude probability of death satisfies $\hat{F}_C(t) = \hat{F}_C(0,t)$. Following Andersen, Borgan, Gill, and Keiding (1993, pp. 290–293) we propose the following estimator for the variance:

$$\widehat{\mathsf{VAR}}(\hat{F}_C(t)) = \int_0^t [\hat{S}_O(u)]^2 [1 - \hat{F}_C(u,t)]^2 \frac{dN(u)}{Y(u)^2}. \tag{8}$$

3. R functions and technical considerations

The three concepts are joined into two main functions:

- `rs.surv`: This function estimates net survival or relative survival ratio. The desired estimator is chosen using the argument `method`:
  - `method = "pohar-perme"`: The net survival estimator with the cumulative hazard given by (6). This method is chosen by default.
  - `method = "ederer1"`: The relative survival ratio estimator given by (4).

  - `method = "hakulinen"`: The correction of the relative survival ratio useful in the presence of informative (covariate-dependent) censoring due to the heterogeneity of potential follow-up times (Hakulinen and Tenkanen 1987). Since this is an ad-hoc correction that does not entirely remove the bias and can introduce additional bias in the presence of non-informative censoring (Rebolj Kodre and Pohar Perme 2013), we do not recommend this method to be used. It is nevertheless included for historical reasons and comparisons.
  - `method = "ederer2"`: Another method included mainly for historical reasons and comparisons, results in biased estimation of net survival (Pohar Perme *et al.* 2012). An age-standardized version of this estimator can have a smaller bias and is more frequently used.

- `cmp.rel`: The function for estimating the crude probability of death from the disease in question $\hat{F}_C(t)$ (7) and the crude probability of death from other causes $\hat{F}_P(t)$.

In terms of computational options available, the `rs.surv` function mimics the `survfit` function of the **survival** package (Therneau 2018) while the `cmp.rel` function follows the `cuminc` function of the **cmprsk** package (Gray 2014).

As in the `survfit` function, we allow two options to calculate the survival function from the cumulative hazard. The `"kaplan-meier"` option uses the formula $\hat{S}(t) = \prod_{(0,t]}\{1 - d\hat{\Lambda}(s)\}$, while the `"fleming-harrington"` method uses the exponential association between the functionals, i.e., $\hat{S}(t) = \exp\{-\hat{\Lambda}(t)\}$. The two options are available for all the methods implemented in `rs.surv`, in case of relative survival ratio, the overall cumulative hazard is given by (5), while the cumulative hazard for the net survival is given by (6).

Several options are also available for the calculation of the confidence intervals – the variance is reported on the cumulative hazard scale and the `conf.type` options for the calculation of the confidence intervals return the `"log-log"`, `"log"` or `"plain"` versions of the confidence intervals.

The `cmp.rel` function allows for less options, as in the `cuminc` function, the observed survival $\hat{S}_O$ in the formula (7) is always calculated using the cumulative product and the variance is reported on the cumulative probability scale with the confidence intervals symmetrical on the same scale.

3.1. Expected number of years lost

An additional parameter that may be of interest in the analysis is the average number of years lost until a certain time point. As presented in Andersen (2013), the integral under each cumulative probability curve until a given time τ can be interpreted as the number of years lost to that cause compared to a cohort where nobody dies before τ. We can thus split the total number of years lost in a certain interval into the number of years lost due to the disease of interest and the number of years lost due to other causes. The values are automatically reported in the output of the `cmp.rel` function as `area`. We limit ourselves to reporting the number of years lost until time point τ to avoid extrapolation beyond the last observation time. This time point can be set with the argument `tau`, the default is the maximum observation time. Note that `tau` does not only affect the calculation of number of years lost, but also the final point until which the curve is calculated – all individuals are censored beyond `tau`.

3.2. Comparison of net survival curves

Recently a new test for comparison of net survival curves has been proposed (Grafféo, Castell, Belot, and Giorgi 2016). It combines the ideas from the PP estimator (Pohar Perme *et al.* 2012) and the log-rank test statistic (Fleming and Harrington 1991). Its properties have been further explored in Pavlič and Pohar Perme (2017). Both the stratified and nonstratified version of the test have been developed and both are included in the function `rs.diff`.

3.3. Net expected sample size

Some authors (Lambert, Dickman, and Rutherford 2015; Dickman, Lambert, Coviello, and Rutherford 2013) report overly large variability when using the PP estimator, particularly when considering long-term net survival. While this may seem like a practical issue with a particular estimator, it is indeed an intrinsic property of the definition of the net survival. Since net survival is defined as the survival in the hypothetical world where individuals can die only of cancer, one cannot estimate it if no data on this world are available, i.e., if all patients of a certain group die of other causes. In other words, it simply does not make sense to estimate 15-year net survival of patients aged 90, since their probability of being still alive at that time even if they do not have the disease is practically 0. Since the overall net survival is the average over all individuals in the sample (2) it is crucial that the estimation is sensible for all individuals in the sample. Therefore, one must either limit the calculation to the follow-up interval in which all patients included have a large enough probability of not yet dying due to other causes or consider only a subgroup of patients for which this is true. By limiting to age groups for which we can expect enough patients to be still alive by the time of interest, we do not throw away data but rather limit the estimation to the subset for which the information is actually available. If we nevertheless wish to estimate long term net survival for all individuals, some parametric assumptions and hence extrapolation of the required information must be made.

As a guideline on what might still be sensible, we provide a function `nessie` that calculates the net expected sample size, i.e., the number of people that are still exposed at a certain time point after the expected deaths due to population reasons are removed. This should provide some insight into the length of the time interval in which it is still sensible to estimate net survival for a given age group and to thus avoid estimation based on very few individuals. Note that the censoring pattern is not included in this calculation, so the expected numbers are often even lower. As an alternative possible guideline, we also report the expected remaining lifetime of a certain age group in the population.

3.4. Relative survival particularities

This subsection covers some important differences between the estimators in the classical and relative survival field, i.e., points where one should be careful and cannot directly use the classical survival analogy. While this subsection can be skipped by the first-time users of the relative survival methodology, it is crucial for a deeper understanding of the estimators in the field.

The estimators are not step functions

The most important difference to note is that while the value of stochastic integrals with

respect to dN (e.g., the first integral in (6)) only jumps at event times, the cumulative population hazard Λ_P is a continuous function. The integral with respect to $d\Lambda_P$ (e.g., the second integral in (6)) is continuously changing between event times, which means that the estimators are not step functions. All estimators of survival shall increase between the event times and jump at event times whereas the estimate of the crude probability of death shall decrease between event times and also jump at event times. This is true for all nonparametric relative survival estimators, though it has, to our knowledge, never been specifically mentioned or cared about in practice. Instead, when reporting the estimated value at a given time point (say 5 years) at which there was no event, the last value is carried forward, though this incurs some bias. The size of this bias depends on the length of the gaps between event times, however, with the large data sets typically occurring in the field, it is often negligible in practice.

Population hazard changes in time

To understand how the population survival in time is calculated in our functions, consider the integral (3). A standard population mortality table typically reports the yearly probabilities split by age, sex and year. More precisely, they report the probability that a person of a certain sex and of age a at the beginning of year y survived until the end of that year. Under the assumption that the hazard was constant within that year, the daily hazards λ_P are calculated for each combination of age, year and sex and included in the tables. When using these tables to get $\lambda_{Pi}(t)$ for an individual i, the value which corresponds to the age and calendar year of person i at time t is considered. This means that the $\lambda_{Pi}(t)$ used for calculations for each individual i changes in time – it starts at the age and year of diagnosis and then changes when the individual either gets one year older or a new calendar year starts. Therefore, $\lambda_{Pi}(t)$ is a step-wise constant function of time that changes twice a year for each individual, the times of jumps are different for each individual. The integral $\Lambda_{Pi}(t)$ is an increasing piecewise linear function.

Controlling how the population hazards are used in the functions

Since the value of $\lambda_{Pi}(t)$ changes at different times for each individual, the actual calculation is made by splitting into small intervals in which $\lambda_{Pi}(t)$ is regarded as constant. In all functions, $d\hat{\Lambda}_P(t)$ is then calculated as $\lambda_P(t) \cdot dt$. By default, the argument `precision` which specifies the length of these intervals is set to 1, which implies daily intervals. In practice, taking daily intervals should suffice for any calculation, since the fact that $\lambda_{Pi}(t)$ is a step-wise constant function is anyway an artefact of the available data. However, the estimated values might change slightly if even narrower intervals are set.

Numerical integration in the PP estimator

In the case of the PP estimator (and the log-rank type test, which follows the same logic), the integration between event times is slightly more complex than in other cases, as the second integral in (6) contains also $S_{Pi}(u)$ which continuously decreases. Therefore, numerical integration is needed – we calculate the integral as the average of the values at the first and last point of the interval times the length of the interval. Again, the default for `precision` is set to 1 day.

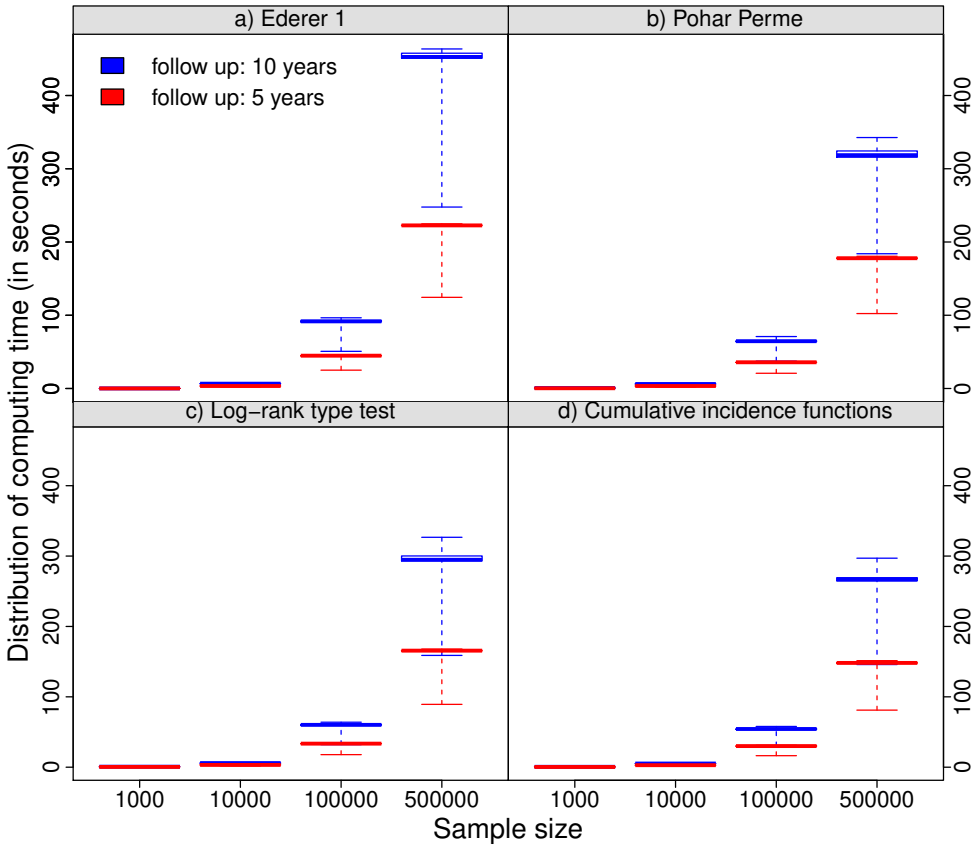

Figure 1: Distribution of computing times for different functions from the package.

Controlling the times at which the results are reported

Note that having more terms in the sum can considerably increase the computational intensity if the event times are few, however, the default precision shall only add few time points with large data sets where the event times already occur almost daily. Estimates at these additional time points are not included in the output to keep the output manageable and to be consistent with the **survival** package where the output includes only the results at observed times. An additional argument `add.times` is then included to ensure correct reporting of the results at pre-given time points that do not equal any of the observed times.

Computational intensity

Since the cancer registry data sets may be very large (more than 100000 patients), the calculation of the estimators and their variances may become computationally very intensive, in particular if performed in short intervals. To speed up the calculation, most functions use subroutines written in C that considerably speed up the process, C subroutines written for the **survival** package (Therneau 2018) are also included. The total processing time depends on the number of individuals and the number of unique event times. To illustrate the computing times of different functions, we performed a small simulation study. The distribution of computing time (in seconds) is presented in Figure 1. The functions are written sufficiently fast that they give results in a few minutes even for samples of 500000 patients.

Simulation details. An exponential model was used for the excess hazard. The follow-up time was either 5 or 10 years – around 30% or 50% of patients had an event in the first or in the second case, respectively. A log-rank type test was used to compare two groups of equal size with the same excess hazard. In each case 100 samples were simulated and the computing times for the different functions were measured.

4. Usage

Ignoring for the moment all the additional options available, the `rs.surv`, the `cmp.rel` and the `rs.diff` functions all have the same basic syntax:

```
rs.surv(formula, data, ratetable)
cmp.rel(formula, data, ratetable)
rs.diff(formula, data, ratetable)
```

The data on the observed cohort are passed through the argument `data`, the mortality table to be used should be specified with the argument `ratetable`. The mortality tables need to be organized as a '`ratetable`' object which is defined in the **survival** package. For all the details on this object see Therneau and Offord (1999); further advice on its usage and purpose-made functions to simplify this work can be found in Pohar and Stare (2007) or Pohar and Stare (2006). While it may be time consuming to organize a table of population mortality hazards when first importing it into R, no further reorganization of this object is needed for each of the **survival** or **relsurv** package functions. Using and comparing different estimators is thus particularly simple in R.

The syntax of the `formula` equals that of the **survival** package.

```
formula = Surv(time, cens) ~ 1
```

The '`Surv`' object contains the follow-up time and the censoring indicator, which equals 1 for a time of death (of any cause) and 0 for the time when a person is lost from follow-up. It is important that the follow-up time is always expressed in days, since the hazards in the '`ratetable`' objects are also expressed in days. The value 1 to the right of the `~` sign indicates that only curves for the entire cohort are required – if one wishes to estimate curves with respect to subgroups formed by a certain variable, that variable (or a sum of several variables) should be written to the right of the `~` sign.

If the demographic covariates by which the mortality tables are split (usually age, sex and calendar year) are not organized or named in the same way in the observed data set on the cohort as they are in the population tables ('`ratetable`' object), they can be properly matched using the argument `rmap`. Note that the calendar year must be in a date format (date, Date and POSIXt are allowed), but the date formats in the `ratetable` and in the data may differ.

Several functions of the package need to transform between days and years, the factor `365.241` is used for this transformation in all the cases. Therefore, whenever this transformation is used with the data, the same factor should be used.

All three functions have methods for printing the output and the first two also have methods for plotting the curves, each of them mimicking the analogous methods in the classical survival

functions. Additionally, the `summary` method from the **survival** package may also be used for printing the `rs.surv` output at specific time points. Note that since this is a **survival** package function, it assumes a step-wise function between event times, the option `add.times` should be used in the `rs.surv` function when we wish to evaluate survival also at specific time points (additional to all observed times). The standard error reported with the `summary` method is the standard error of the net survival curve, while the confidence intervals are calculated using the method specified with the `conf.type` argument in `rs.surv`.

The `summary` method also prints the output of `cmp.rel` at specific time points, again the `add.times` option in the `cmp.rel` function ensures that the last value is not simply carried forward and that the output is actually evaluated at that time.

By default, the `plot` method plots the curve at event and censoring times only (and, if specified, at times added by `add.times`), a step curve is drawn in between. This is only an approximation of the curve, for more accuracy between these points, the argument `all.times` should be set to `TRUE`, which shall return a more ragged but more exact curve (this option will plot the curve at all times at which it was estimated, i.e., also at times determined by the argument `precision`).

Other functions that can be useful in the analysis are also included in the **relsurv** package. The functions `transrate`, `transrate.hmd`, `transrate.hld` and `joinrate` may be useful when organizing the mortality tables and `rsadd` can be used to fit the Estève additive model (Estève, Benhamou, Croasdale, and Raymond 1990) and thus compare the curves within subgroups. The functions were described in Pohar and Stare (2006) and Pohar and Stare (2007).

5. Example

To illustrate the usage of the functions from the **relsurv** package we will use a subset of the data set `colrec` which is included in the package. This data set consists of 5971 patients diagnosed with colon or rectum cancer between January 1st, 1994 and December 31st, 2000. It has been provided by the Cancer Registry of Slovenia and analyzed in Zadnik, Primic Žakelj, and Krajc (2012) and Zadnik, Žagar, and Primic Žakelj (2016). The age, time and date of diagnosis variables are randomly perturbed to make the identification of patients impossible.

The goal of our illustrative example is to compare 10-year survival of patients diagnosed with colon cancer from January 1st, 1994 to December 31st, 1995 to survival of those diagnosed from January 1st, 1999 to December 31st, 2000. The subsets were chosen only as an example and since the data are perturbed to some extent, no medical conclusions should be made based on these results. We nevertheless attempt some interpretation of the results to help the user in this integral part of the analysis. Our analysis shall be performed in the following steps:

- *Forming the data set*: Choosing the subset of patients diagnosed with colon cancer during the two periods; censor them after ten years and add a variable that indicates the period of diagnosis.
- *Importing the* '`ratetable`' *object*: Import and check the table of event rates ('`ratetable`' object) if it is already available; construct it otherwise.
- *Matching the variables*: Match the data set to the '`ratetable`' object.

- *Estimation of relative survival ratio* for the two periods of diagnosis.
- *Limiting the data set*: Limit the analysis to the subgroups of patients for which it is sensible to estimate net survival after ten years.
- *Estimation and comparison of net survival* for the two periods of diagnosis.
- *Estimation of crude probability of death* for the two periods of diagnosis.

Before we proceed we have to load the **relsurv** package.

```
R> library("relsurv")
```

5.1. Forming the data set

Below are the first three lines of the `colrec` data set.

```
R> colrec[1:3, ]
```

```
  sex   age  diag time stat stage   site
1   1 23004 12656   16    0     1 rectum
2   2 12082 13388  504    0     3 rectum
3   1 24277 12711   22    0     3  colon
```

The crucial variables for the relative survival analysis are observed time (`time`) and status (`stat`), gender (`sex`), age at diagnosis (`age`) and date of diagnosis (`diag`). Additionally, the variables `stage` and `site` are included. Gender is coded as `1` for male and `2` for female, `age` and `time` are given in days and `diag` is in date format (days since January 1st, 1960). For our example we choose only two subgroups of patients. To this end, we form an additional variable `d.int` that indicates whether the patient was diagnosed during the first or the second period.

```
R> d1 <- subset(colrec, site == "colon" & diag >= as.date("1Jan1994") &
+    diag <= as.date("31Dec1995"))
R> d1$d.int <- 1
R> d2 <- subset(colrec, site == "colon" & diag >= as.date("1Jan1999") &
+    diag <= as.date("31Dec2000"))
R> d2$d.int <- 2
R> d <- rbind(d1, d2)
```

Since we are interested in 10-year survival, we censor all patients that were still alive after ten years.

```
R> ind <- which(d$time > 365.241 * 10)
R> d$time[ind] <- 365.241 * 10
R> d$stat[ind] <- 0
```

This data set consists of 2003 patients where 883 were diagnosed during the first period and 1120 during the second.

Further notes: The steps described above may not be needed when one wants to analyze one's own data, but they are included anyway for the sake of reproducibility of this example.

5.2. Importing the 'ratetable' object

Since our data set is from the Cancer Registry of Slovenia, we have to use the 'ratetable' object for Slovenia. It is included in the package. It has three dimensions:

```
R> attributes(slopop)$dimid

[1] "age"  "year" "sex"
```

and contains hazards for each combination of covariates from mortality tables. It is thus a tridimensional array. We can look at the hazards for, say, 50 and 70 year old individuals in 1990 and 2000 by using the following line of code.

```
R> slopop[c("50", "70"), c("1990", "2000"), ]

Rate table with dimension(s): age year sex
, , sex = male

    year
age          1990         2000
  50 2.735107e-05 1.537543e-05
  70 1.324940e-04 1.225977e-04

, , sex = female

    year
age          1990         2000
  50 1.036894e-05 8.500730e-06
  70 6.903729e-05 5.377721e-05
```

Note that the hazards are expressed per day, hence the small values. As expected, the hazard is higher for males, older individuals and those who lived earlier. Once the 'ratetable' object is constructed, it can be used with any function from the **relsurv** package without further changes.

Further notes: For other countries such an object may not be available and has to be constructed first. The **relsurv** package includes the following functions to simplify this step: `transrate`, `transrate.hld`, `transrate.hmd` and `joinrate`. The most straightforward to use is the function `transrate.hmd`, which transforms the tables that can be downloaded from the web site Human Mortality Database (HMD, http://www.mortality.org/) to an object of type 'ratetable'. For example, to construct a 'ratetable' object for Slovenia, one should download the yearly "period life tables" (files `mltper_1x1.txt` and `fltper_1x1.txt` for males and females respectively) and use the following code.

```
R> slotab <- transrate.hmd(male = "mltper_1x1.txt",
+    female = "fltper_1x1.txt")
```

5.3. Matching the variables

Having imported the population mortality tables into the format 'ratetable', we now have to match the observed data and the population tables. We have seen that the Slovene 'ratetable' object slopop has dimensions age, year and sex, so the same three variables must exist also in the observed data set. If the names and the format of the variables are equal in both data sets, no further work has to be done, otherwise, one can take care of the matching via the argument rmap in each function call.

In our case, the format of the variables matches (our age is in days, the diagnosis year is in date format), but the names are not the same, we therefore write:

```
rmap = list(age = age, sex = sex, year = diag)
```

Further notes: If age was reported in years and not in days (in a variable named agey), the argument rmap should be

```
rmap = list(age = agey * 365.241, sex = sex, year = diag)
```

5.4. Estimation of relative survival ratio

To estimate the relative survival ratio, we use the function rs.surv with the argument method specified as "ederer1". We estimate it with respect to the variable d.int, which denotes the period in which the patient was diagnosed – this variable is included in the formula described in the previous subsection. We compare the observed cohort to the Slovene population tables and hence set the ratetable argument to slopop. The argument add.times is used to specify that the curve should be evaluated at five and ten years (see Section 4 for details).

```
R> fit_rsr <- rs.surv(Surv(time, stat) ~ d.int,
+    data = d, ratetable = slopop, method = "ederer1",
+    add.times = c(5, 10) * 365.241,
+    rmap = list(age = age, sex = sex, year = diag))
```

Methods such as summary and plot can be used to explore the results. To print the estimated values of the relative survival ratio at five and ten years, we write:

```
R> summary(fit_rsr, times = c(5, 10) * 365.241)
```

```
Call: rs.surv(formula = Surv(time, stat) ~ d.int, data = d,
    ratetable = slopop, method = "ederer1", add.times = c(5, 10) *
    365.241, rmap = list(age = age, sex = sex, year = diag))

                d.int=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
```

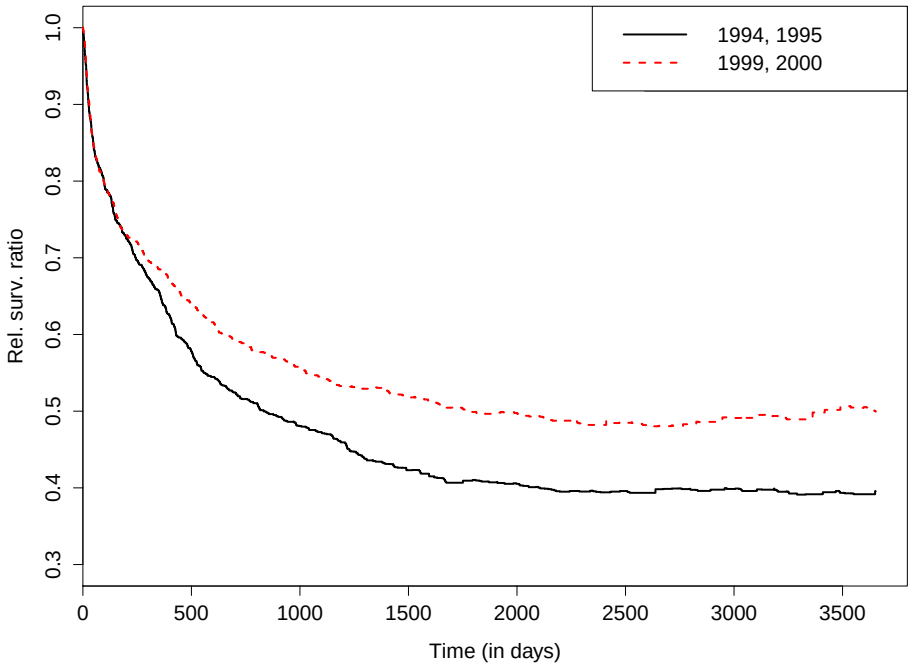

Figure 2: Relative survival ratio for patients diagnosed in the first period (black) and in the second period (red).

```
1826     287      594    0.409  0.0198        0.372        0.450
3652     216       71    0.396  0.0234        0.353        0.444

               d.int=2
time n.risk n.event survival std.err lower 95% CI upper 95% CI
1826    441     679    0.497  0.0184        0.462        0.535
3652    332     109    0.493  0.0227        0.451        0.540
```

The relative survival ratio for those diagnosed during the first period is lower compared to the relative survival ratio of those diagnosed during the second period. This means that even though the population mortality has improved between the two periods, the observed survival of the patients has improved even more, thus increasing the relative survival ratio. The same can be seen in Figure 2.

5.5. Limiting the data set

Since we are interested in estimating 10-year net survival, we have to limit ourselves to those patients for which such an estimate is sensible, i.e., their probability not to have died due to other causes in that period is high enough (see Section 3.3 for details). The function `nessie` reports the number of patients we can expect to remain at risk after a certain time if our patients died due to population hazards only. As this is a guideline only, the choice of age groups in which we do the calculation is arbitrary, we choose 5-year age intervals.

```
R> breaks <- c(0, seq(from = 45, to = 90, by = 5), Inf)
R> d$agegr <- cut(d$age / 365.241, breaks)
```

We call the function with the same syntax as in the previous section:

```
R> nessie(Surv(time, stat) ~ d.int + agegr,
+    data = d[d$age / 365.241 > 70,], ratetable = slopop,
+    times = seq(0, 10, 2), rmap = list(age = age, sex = sex, year = diag))
```

The net expected sample size is estimated with respect to two different time periods (variable `d.int`) and with respect to different age groups (variable `agegr`). The first variable is included because we wish to produce a separate net survival estimate in each of these two calendar periods and the second one is included to give us some insight on what is the oldest age group for which it is still sensible to estimate 10-year net survival. Since only older patients can be problematic, we limit ourselves to individuals above 70. The `times` argument specifies that the estimation is required in two-year long intervals.

```
                        0     2     4     6     8    10 c.exp.surv
d.int=1,agegr(70,75]  150 137.2 123.8 110.0  95.7  80.7       11.0
d.int=1,agegr(75,80]   73  63.5  53.8  44.3  35.2  26.7        8.6
d.int=1,agegr(80,85]   87  68.4  51.3  36.8  25.0  15.8        5.9
d.int=1,agegr(85,90]   40  26.8  16.9  10.0   5.6   2.8        4.1
d.int=1,agegr(90,Inf]   4   2.2   1.1   0.5   0.2   0.0        2.9
d.int=2,agegr(70,75]  207 190.3 172.9 154.5 134.8 114.7       11.8
d.int=2,agegr(75,80]  162 142.8 122.6 101.2  74.2  51.8        8.4
d.int=2,agegr(80,85]   62  50.0  38.5  28.3  20.8  14.5        6.5
d.int=2,agegr(85,90]   65  45.6  29.9  18.0   9.3   4.3        4.3
d.int=2,agegr(90,Inf]  21  11.3   5.3   1.9   0.5   0.1        2.7
```

As we can see, the net expected sample sizes after ten years in the first time period are only 15.8, 2.8 and 0.0 for the oldest three age groups. Also, the expected life time for those between 80 and 85 years old is only 5.9 years. Similar estimates can be seen in the second time period. In our data set, we can expect even considerably less patients, since the patients shall also die of cancer. Therefore, following the above table, we focus on patients aged 80 years or less at the time of diagnosis.

```
R> d2 <- d[d$age < 80 * 365.241, ]
```

This data set consists of 1724 patients aged 80 or less (752 patients diagnosed in the first and 972 in the second period) and it will be used in the analysis of net survival.

5.6. Estimation and comparison of net survival

To estimate net survival, the function `rs.surv` is used with the argument `method` set to `"pohar-perme"`. As before, estimation is performed with respect to the variable `d.int` and argument `add.times` is used as we shall require the estimates to be reported at 5 and 10 years.

```
R> fit_net <- rs.surv(Surv(time, stat) ~ d.int, data = d2,
+    ratetable = slopop, method = "pohar-perme", add.times = c(5, 10) *
+    365.241, rmap = list(age = age, sex = sex, year = diag))
```

Again, we consider the estimated net survival at five and ten years with the method `summary`.

```
R> summary(fit_net, times = c(5, 10) * 365.241)

Call: rs.surv(formula = Surv(time, stat) ~ d.int, data = d2,
    ratetable = slopop, method = "pohar-perme", add.times = c(5, 10) *
    365.241, rmap = list(age = age, sex = sex, year = diag))

               d.int=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 1826    269     482    0.414  0.0204        0.376        0.456
 3652    212      57    0.396  0.0244        0.351        0.447

               d.int=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 1826    427     545    0.509  0.0187        0.474        0.547
 3652    328      99    0.497  0.0247        0.451        0.548
```

Net survival is higher for the patients diagnosed during the second period and the differences between the periods are similar both at five and ten years. The values imply that in a hypothetical world, where the patients would be exposed to cancer hazard only, the 5-year survival would be 0.41 and 0.51 for the two periods, respectively. The estimated net survival then stays practically equal for the next five years indicating that the hazard of dying due to cancer is practically 0 in that interval.

Having estimated net survival, we have made the two periods directly comparable even if the population mortality has considerably changed in between. The better survival in the second period can be thus attributed to the lowered cancer specific hazard. The only other cause for this difference could be in the different covariate distribution of the patients in the second period (e.g., younger patients, earlier stage, less smoking) – this can then be further investigated using regression modeling (to this end the function `rsadd` can be used, see Pohar and Stare 2006 for details).

Figure 3 presents the estimated net survival of the patients diagnosed in each time period, we can use the log-rank type test to test whether the net survival is significantly different for patients diagnosed in different time periods. To this end, we use the function `rs.diff`.

```
R> rs.diff(Surv(time, stat) ~ d.int, data = d2, ratetable = slopop,
+    rmap = list(age = age, sex = sex, year = diag))

Value of test statistic: 9.254295
Degrees of freedom: 1
P value: 0.002349437
```

Results include the value of the test statistic, the number of degrees of freedom and the p value. As expected from Figure 3 we reject the null hypothesis of equal net survival in the two periods. Using the same function, we can also consider the stratified log-rank test, e.g., test whether the differences persist within different age groups. We use the variable `agegr` to form the strata.

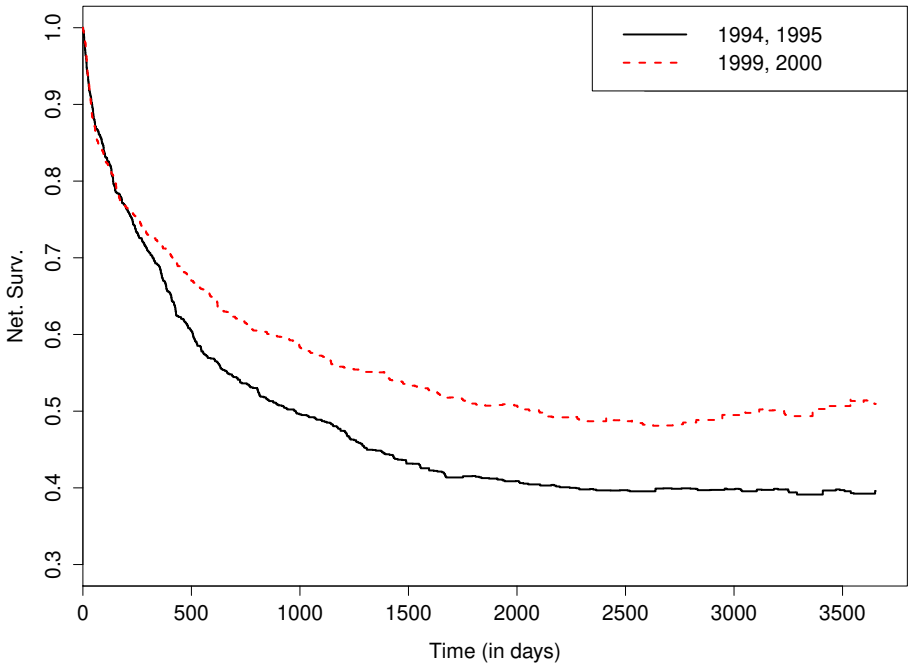

Figure 3: Net survival for patients diagnosed in the first period (black) and in the second period (red).

```
R> rs.diff(Surv(time, stat) ~ d.int + strata(agegr), data = d2,
+    ratetable = slopop, rmap = list(age = age, sex = sex, year = diag))

Value of test statistic: 10.36237
Degrees of freedom: 1
P value: 0.0012861
```

The value of the test statistic has slightly increased. This implies that the difference between net survival in different periods is even larger within the age groups.

Further notes: Function `rs.diff` also has an option `precision` which is by default set to `1`. This value can be decreased to allow even more accurate calculations or increased to allow faster calculations.

```
R> rs.diff(Surv(time, stat) ~ d.int, data = d2, ratetable = slopop,
+    precision = 0.1, rmap = list(age = age, sex = sex, year = diag))

Value of test statistic: 9.253211
Degrees of freedom: 1
P value: 0.002350829
```

Comparing this result with the first one above, we can notice that the increased precision changed the results only minimally. This is in line with our experience, which shows that precision lower than 1 day is practically never needed. Since our data set is rather large, the gaps between event and censoring times are rather small (median gap is 2 days), therefore,

increasing the argument precision also does not change the result (the value of the test statistic becomes equal to 9.29). However, if the gaps between the event and censoring times were larger, setting the precision to smaller intervals is crucial for exact calculation even if it slows down the function's performance.

When considering the log-rank test with less than ten events in any of the groups, the function gives a warning.

5.7. Estimation of crude probability of death

We finally turn to the estimation of the crude probability of death in the two diagnosis periods. We use the `cmp.rel` function for this purpose.

```
R> cmp_fit <- cmp.rel(Surv(time, stat) ~ d.int, data = d,
+    ratetable = slopop, rmap = list(age = age, sex = sex, year = diag))
```

The results of this function can be viewed with the function `summary`. It has four arguments. The first one is a '`cmp.rel`' object, i.e., the output of the function `cmp.rel`, e.g., the object `cmp_fit` in our case. The second argument `times` is used to specify the time points at which the estimates are required, the third argument specifies the units in which the `times` are given, the default is `365.241` and represents years, since we wish a report at 5 and 10 years, the `scale` is set to `365.241` and is included just for the sake of completeness. The last argument is used to specify whether the area under the curve should be printed out.

```
R> summary(cmp_fit, times = c(5, 10), scale = 365.241, area = TRUE)

$`est`
                             5        10
causeSpec d.int=1  0.57954704 0.5980827
population d.int=1 0.09463701 0.1567039
causeSpec d.int=2  0.50712920 0.5237839
population d.int=2 0.09912080 0.1797875

$var
                               5           10
causeSpec d.int=1  3.308489e-04 3.862212e-04
population d.int=1 7.433284e-06 3.216670e-05
causeSpec d.int=2  2.763276e-04 3.408826e-04
population d.int=2 5.279948e-06 2.784507e-05

$area
                    Area at tau = 10
causeSpec d.int=1          5.2602207
population d.int=1         0.9032382
causeSpec d.int=2          4.6132087
population d.int=2         0.9771146
```

The output contains the estimates of cause-specific and population mortality and variances of these estimates at several time points for both groups defined by the variable at the right

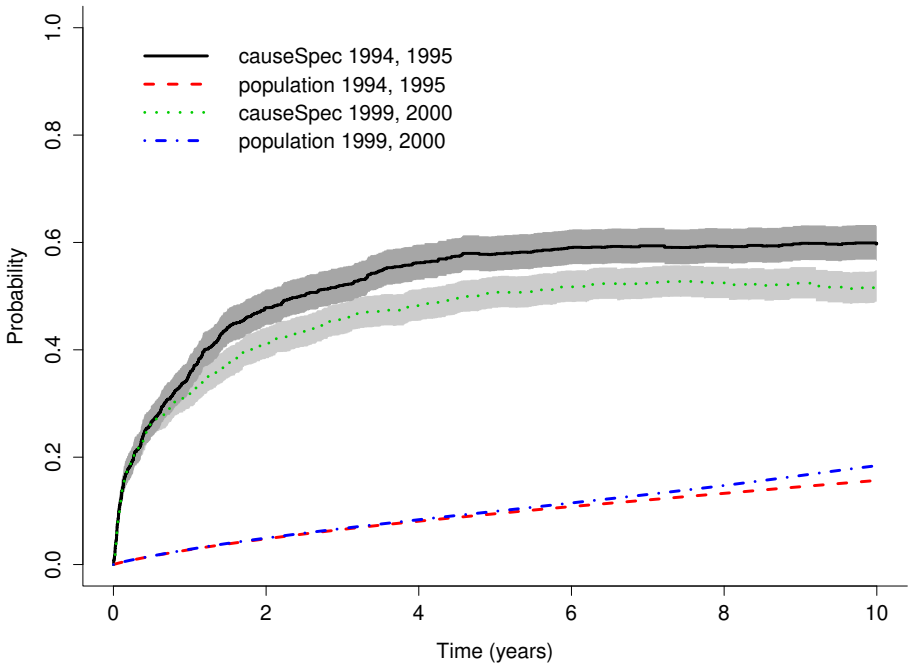

Figure 4: Crude (cause-specific) probability of death curves with confidence intervals and other cause (population) mortality curves for patients diagnosed in the two periods.

hand side of `~` in the `formula` part. It also includes the area under the curve up to time `tau`, which is by default the maximum observed time (ten years in our example, can be set otherwise in the `tau` argument of function `cmp.rel`). Patients diagnosed in the first period have approximately 0.1 higher probability of dying due to the disease at five and ten years than patients in the second period. On the contrary, the probability of dying from other causes is slightly higher in the second period. This can probably be attributed to the fact that fewer patients die from cancer, all the observed results could also be a consequence of the different distribution of covariates in the second period. The theory for exploring this directly via regression models has not been introduced yet in the relative survival field.

The area under the curve tells us that patients diagnosed during the first period have lost approximately 5.3 years due to cancer in the 10-year period, whereas the patients diagnosed in the second period lost 4.6 years. For comparison, the years lost in the same time due to other causes were much fewer – slightly below one year in both periods.

These results can also be presented graphically using the `plot` method.

```
R> plot(cmp_fit, col = 1:4, lwd = 3, xscale = 365.241,
+     xlab = "Time (years)", conf.int = c(3, 1))
```

We have provided several arguments to make this plot more readable; the result is given in Figure 4. The `xscale` puts the scale of ordinal axis into years instead of the default (`1`), which is days. By default, all estimated cumulative incidence curves are plotted, this could be changed with the argument `curves` (the default is `1:4`, i.e., all curves, see the output of `summary` for the order of curves and their total number). The same is true also for the confidence intervals – we choose to plot the confidence intervals for cancer specific curves only

(first and third curve in our case). Notice that we can specify the order in which confidence intervals are to be plotted to emphasize how they overlap (Figure 4).

Further notes: Function `cmp.rel` prints warnings when it has issues with the calculation of confidence intervals for the crude probability of death. When the estimated variance is negative, the square root of variance cannot be evaluated and the standard deviation cannot be obtained. This will often happen in the early intervals and sometimes towards the end of follow-up. The graph can be used to further evaluate the importance of this warning (intervals with the negative estimated variance shall be missing). The function `cmp.rel` also has arguments `add.times` and `precision` that play the same role as in the function `rs.surv`. When one wants to estimate crude cause-specific probability of death in a shorter time interval or the areas under these curves are of interest up to a specific time point the argument `tau` can be used. By default it is set to the maximum observed time. If we are interested in the areas under the curves at five years, we can set it to `5 * 365.241`.

```
R> cmp_fit2 <- cmp.rel(Surv(time, stat) ~ d.int, data = d,
+    ratetable = slopop, tau = 5 * 365.241,
+    rmap = list(age = age, sex = sex, year = diag))
```

The '`cmp.rel`' object is a list, where the length matches the number of estimated curves plus one – the last element is the value of the argument `tau`. The output can also be read directly, without using the `summary` method, e.g., areas under the crude cause-specific probability of death curves in both time intervals can be obtained in the following way:

```
R> cmp_fit2[[1]]$area

[1] 2.29866

R> cmp_fit2[[3]]$area

[1] 2.012484
```

We can see that patients diagnosed during the first period have lost around 2.3 years due to cause-specific reasons in five years and patients diagnosed in the second period have lost around 2 years due to the cause-specific reasons in a five years time. In a similar fashion the values of estimators, variances and lower or upper boundaries of confidence intervals can be obtained.

The results of the function `cmp.rel`, i.e., a '`cmp.rel`' object, can be also printed with the method `print`, which chooses the time points for output by itself. The `summary` method is used as an alternative with more user control.

6. Discussion and conclusions

Several new advances have been made in the field of relative survival over the past decade. Among them the theoretical clarification of the different measures and the new proposal for net survival estimator were a key step (Pohar Perme *et al.* 2012). Furthermore, the need for

estimating crude probability of death has been emphasized (Eloranta, Adolfsson, Lambert, Stattin, Akre, Andersson, and Dickman 2013).

A substantial gap between the theory available and the methods in use can be observed, with the estimators that have been shown not to be consistent (e.g., using Ederer II method for estimation of net survival) still being frequently used.

By making the new developments available in a user friendly software, we hope to decrease the gap between the theory and practice – we ensure that the methods can be more directly used and also that the properties of the methods can be further studied. Some of the proposed methodology requires only ad-hoc changes of the existing functions (e.g., age-standardized Ederer II). The focus of this paper is on the two estimators, where the algorithm is rather complex. Both the PP estimator of net survival and the continuous-time estimator of crude probability of death require the population mortality hazard to be known for each individual at all times while still alive, thus making the matching of the observed data and the population tables a nuisance that prevents even the more enthusiastic users from programming the functions by themselves. We explain the specifics of the relative survival estimators which make any simplifications of these estimators biased. In particular, these specifics help understand why the estimator shall be biased when only discretely recorded times of events are available (for example only the number of events per month). While some ad-hoc methods for accounting for this problem have been proposed (Seppä, Hakulinen, and Pokhrel 2015), this requires some future work in terms of theory and software development.

When comparing our package to other software packages, Stata is the only one with the same extent of methodology available, while others like SAS (SAS Institute Inc. 2015) and SEER*Stat (Surveillance Research Program 2016) are still lagging behind. Three commands for net survival estimation exist in Stata (`stns`, Clerc-Urmès, Grzebyk, and Hédelin 2014; `strs`, Dickman and Coviello 2015; `stnet`, Coviello, Dickman, Seppä, and Pokhrel 2015) and in SEER*Stat the PP estimator is available as of version 8.3.1. While the command `stns` uses the same algorithm as our function `rs.surv` in R, the commands `strs` and `stnet` use a life-table approach based on the idea of inverse weighting from the PP estimator. Since this approach can produce a non-negligible bias when the intervals between events are too wide, some further work has been done to account for that (Seppä *et al.* 2015). The only current difference between the `rs.surv` in R and `stns` in Stata is the fact that `stns` calculates the estimates only at observed times and assumes a step-function in between – when the gaps between the event times are small, the results of the two functions are practically the same. Both the `strs` and `stnet` commands also provide the Ederer II estimator of relative survival ratio and the first one also includes commands that have traditionally been used for net survival estimation (Hakulinen, Ederer I). For interval data, a nonparametric estimation of the crude probability of death function (Cronin and Feuer 2000) is also available in Stata command `strs`. The output of the `stns` function offers all the parts needed for the estimation of crude probability of death (but not its variance). On the other hand, a considerable amount of work has been done in Stata in terms of predicting crude probability curves based on a flexible parametric model (Royston and Lambert 2011).

Though the focus of this paper is on two nonparametric methods, package **relsurv** includes all the necessary tools for a high quality relative survival analysis, from functions for importing the population tables (which try simplifying this, typically most time-consuming part of any relative survival analysis) to regression modeling. The paper also describes the most important recent inclusions, for which our package is still the only existing software package,

but which we believe may be useful in any quality nonparametric analysis: the log-rank type test for comparison of net survival curves, the calculation of the area below the crude probability curves which can be interpreted as the number of years lost by the patient and the calculation of the net expected sample size which can provide a guideline for sensible estimation of net survival.

Acknowledgments

Both authors are employed at the Institute for Biostatistics and Medical Informatics, Faculty of Medicine, University of Ljubljana. Klemen Pavlič is a young researcher funded by the Slovenian Research Agency. This research has been conducted as a part of the project "Methods of estimation of key indicators in population cancer survival" (J3-7272) funded by the Slovenian Research Agency.

The authors are grateful to the Cancer Registry of Slovenia for providing the data.

References

Andersen PK (2013). "Decomposition of Number of Life Years Lost According to Causes of Death." *Statistics in Medicine*, **32**(30), 5278–5285. doi:10.1002/sim.5903.

Andersen PK, Borgan O, Gill RD, Keiding N (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York. doi:10.1007/978-1-4612-4348-9.

Charvat H, Belot A (2018). ***mexhaz**: Mixed Effect Excess Hazard Models*. R package version 1.5, URL https://CRAN.R-project.org/package=mexhaz.

Clements M, Liu XR (2018). ***rstpm2**: Generalized Survival Models*. R package version 1.4.2, URL https://CRAN.R-project.org/package=rstpm2.

Clerc-Urmès I, Grzebyk M, Hédelin G (2014). "Net Survival Estimation with `stns`." *The Stata Journal*, **14**(1), 87–102.

Coviello E, Dickman PW, Seppä K, Pokhrel A (2015). "Estimating Net Survival Using a Life-Table Approach." *The Stata Journal*, **15**(1), 173–185.

Cronin KA, Feuer EJ (2000). "Cumulative Cause-Specific Mortality for Cancer Patients in the Presence of Other Causes: A Crude Analogue of Relative Survival." *Statistics in Medicine*, **19**(13), 1729–1740. doi:10.1002/1097-0258(20000715)19:13<1729::aid-sim484>3.0.co;2-9.

Dickman PW, Coviello E (2015). "Estimating and Modeling Relative Survival." *The Stata Journal*, **15**(1), 186–215.

Dickman PW, Lambert PC, Coviello E, Rutherford MJ (2013). "Estimating Net Survival in Population-Based Cancer Studies." *International Journal of Cancer*, **133**, 519–521. doi:10.1002/ijc.28041. Letter to the Editor.

Ederer F, Axtell LM, Cutler SJ (1961). "The Relative Survival Rate: A Statistical Methodology." *National Cancer Institute Monograph*, **6**, 101–121.

Eloranta S, Adolfsson J, Lambert PC, Stattin P, Akre O, Andersson TM, Dickman PW (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**(3), 505–515. doi:10.1007/s10552-012-0141-5.

Estève J, Benhamou E, Croasdale M, Raymond M (1990). "Relative Survival and the Estimation of Net Survival: Elements for Further Discussion." *Statistics in Medicine*, **9**(5), 529–538. doi:10.1002/sim.4780090506.

Fleming TR, Harrington DP (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons.

Grafféo N, Castell F, Belot A, Giorgi R (2016). "A Log-Rank Type Test to Compare Net Survival Distributions." *Biometrics*, **72**(3), 760–769. doi:10.1111/biom.12477.

Gray B (2014). ***cmprsk**: Subdistribution Analysis of Competing Risks*. R package version 2.2-7, URL https://CRAN.R-project.org/package=cmprsk.

Hakulinen T, Seppä K, Lambert PC (2011). "Choosing the Relative Survival Method for Cancer Survival Estimation." *European Journal of Cancer*, **47**(14), 2202–2210. doi:10.1016/j.ejca.2011.03.011.

Hakulinen T, Tenkanen L (1987). "Regression Analysis of Relative Survival Rates." *Journal of the Royal Statistical Society C*, **36**(3), 309–317. doi:10.2307/2347789.

Lambert PC, Dickman PW, Nelson CP, Royston P (2010). "Estimating the Crude Probability of Death Due to Cancer and Other Causes Using Relative Survival Models." *Statistics in Medicine*, **29**(7–8), 885–895. doi:10.1002/sim.3762.

Lambert PC, Dickman PW, Rutherford MJ (2015). "Comparison of Different Approaches to Estimating Age Standardized Net Survival." *BMC Medical Research Methodology*, **15**(64). doi:10.1186/s12874-015-0057-3.

Pavlič K, Pohar Perme M (2017). "On Comparison of Net Survival Curves." *BMC Medical Research Methodology*, **17**, 79. doi:10.1186/s12874-017-0351-3.

Pavlič K, Pohar Perme M (2018). "Using Pseudo-Observations for Estimation in Relative Survival." *Biostatistics*. doi:10.1093/biostatistics/kxy008.

Pohar M, Stare J (2006). "Relative Survival Analysis in R." *Computer Methods and Programs in Biomedicine*, **81**(3), 272–278. doi:10.1016/j.cmpb.2006.01.004.

Pohar M, Stare J (2007). "Making Relative Survival Analysis Relatively Easy." *Computers in Biology and Medicine*, **37**(12), 1741–1749. doi:10.1016/j.compbiomed.2007.04.010.

Pohar Perme M (2018). ***relsurv**: Relative Survival*. R package version 2.2-3, URL https://CRAN.R-project.org/package=relsurv.

Pohar Perme M, Estève J, Rachet B (2016). "Analysing Population-Based Cancer Survival – Settling the Controversies." *BMC Cancer*, **16**(933), 1–8. doi:10.1186/s12885-016-2967-9.

Pohar Perme M, Stare J, Estève J (2012). "On Estimation in Relative Survival." *Biometrics*, **68**(1), 113–120. doi:10.1111/j.1541-0420.2011.01640.x.

Pokhrel A, Hakulinen T (2009). "Age-Standardisation of Relative Survival Ratios of Cancer Patients in a Comparison Between Countries, Genders and Time Periods." *European Journal of Cancer*, **45**(4), 642–647. doi:10.1016/j.ejca.2008.10.034.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https//www.R-project.org/.

Rebolj Kodre A, Pohar Perme M (2013). "Informative Censoring in Relative Survival." *Statistics in Medicine*, **32**(27), 4791–4802. doi:10.1002/sim.5877.

Royston P, Lambert PC (2011). *Flexible Parametric Survival Analysis Using Stata: Beyond the Cox Model*. Stata Press, College Station. URL http://www.stata-press.com/books/flexible-parametric-survival-analysis-stata/.

SAS Institute Inc (2015). *SAS 9.4 SQL Procedure User's Guide, Third Edition*. Cary. URL http://www.sas.com/.

Seppä K, Hakulinen T, Pokhrel A (2015). "Choosing the Net Survival Method for Cancer Survival Estimation." *European Journal of Cancer*, **51**(9), 1123–1129. doi:10.1016/j.ejca.2013.09.019.

StataCorp (2015). *Stata Statistical Software: Release 14*. College Station. URL http://www.stata.com/.

Surveillance Research Program (2016). *National Cancer Institute SEER*Stat Software Version 8.3.2*. URL http://seer.cancer.gov/seerstat/.

Therneau T (2018). ***survival**: A Package for Survival Analysis in S*. R package version 2.42-6, URL https://CRAN.R-project.org/package=survival.

Therneau T, Offord J (1999). "Expected Survival Based on Hazard Rates (Update)." *Technical Report 63*, Section of Biostatistics, Mayo Clinic.

Zadnik V, Primic Žakelj M, Krajc M (2012). "Cancer Burden in Slovenia in Comparison with the Burden in Other European Countries." *Zdravniški Vestnik*, **81**, 407–412.

Zadnik V, Žagar T, Primic Žakelj M (2016). "Cancer Patients' Survival: Standard Calculation Methods and Some Considerations Regarding Their Interpretation." *Zdravstveno Varstvo*, **55**, 134–141.

Affiliation:

Maja Pohar Perme
Institute for Biostatistics and Medical Informatics
University of Ljubljana, Faculty of Medicine
Vrazov trg 2
1000 Ljubljana, Slovenia
E-mail: maja.pohar@mf.uni-lj.si
URL: http://ibmi.mf.uni-lj.si/en

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2018, Volume 87, Issue 8 *Submitted:* 2016-05-04
doi:10.18637/jss.v087.i08 *Accepted:* 2017-10-23