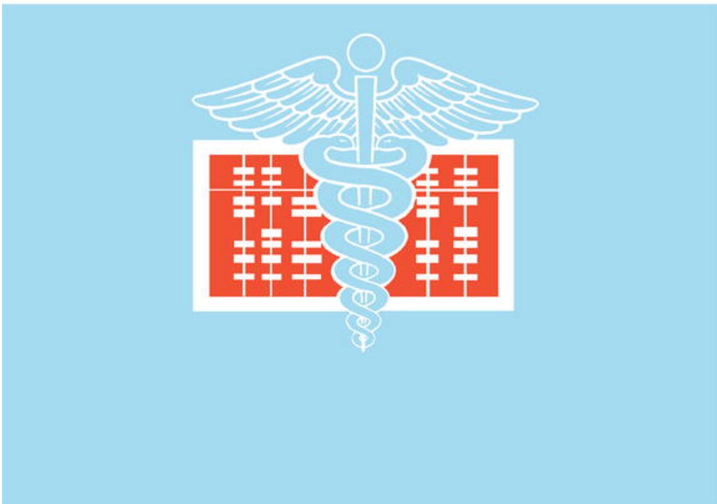
Volume 37 Issue 12 December 2007 ISSN 0010-4825

ELSEVIER

Computers in Biology and Medicine

An International Journal



ELSEVIER

Computers in Biology and Medicine 37 (2007) 1741 – 1749

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Making relative survival analysis relatively easy

Maja Pohar*, Janez Stare

Department of Medical Informatics, University of Ljubljana, Vrazov trg 2, SI-1000 Ljubljana, Slovenia

Received 3 October 2006; accepted 25 April 2007

Abstract

In survival analysis we are interested in time from the beginning of an observation until certain event (death, relapse, etc.). We assume that the final event is well defined, so that we are never in doubt whether the final event has occurred or not. In practice this is not always true. If we are interested in cause-specific deaths, then it may sometimes be difficult or even impossible to establish the cause of death, or there may be different causes of death, making it impossible to assign death to just one cause. Suicides of terminal cancer patients are a typical example. In such cases, standard survival techniques cannot be used for estimation of mortality due to a certain cause. The cure to the problem are relative survival techniques which compare the survival experience in a study cohort to the one expected should they follow the background population mortality rates. This enables the estimation of the proportion of deaths due to a certain cause. In this paper, we briefly review some of the techniques to model relative survival, and outline a new fitting method for the additive model, which solves the problem of dependency of the parameter estimation on the assumption about the baseline excess hazard. We then direct the reader's attention to our R package `relsurv` that provides functions for easy and flexible fitting of all the commonly used relative survival regression models. The basic features of the package have been described in detail elsewhere, but here we additionally explain the usage of the new fitting method and the interface for using population mortality data freely available on the Internet. The combination of the package and the data sets provides a powerful informational tool in the hands of a skilled statistician/informatician.

Keywords: Relative survival; R software; Regression; Population tables

1. Motivation

If a person with an incurable disease commits suicide, the cause of death written in the death certificate will be suicide. And if there were many such cases, the mortality statistics would show much lower proportion of deaths due to the disease in question than it really should. And while suicides are just an obvious, more or less hypothetical, example, it is less well known that it is often difficult or even impossible to select among different possible causes of death or assign a certain cause at all. People with a certain condition (e.g. diabetes, high blood pressure, etc.) may die of natural causes, but it is quite possible that they would have lived longer without that condition. In such cases we need methods of relative survival to estimate the proportion of people dying due to a certain cause. These methods are widely used in cancer registries, but almost never in other areas of medicine.

The goal of this paper is to bring the methods of relative survival to a wider, possibly less statistical, audience, by giving an overview of the existing methods, outline some new methods, describe new possibilities of acquiring population data, and present a software package that includes functions for easy and flexible fitting of all the commonly used relative survival regression models.

2. Introduction

In survival analysis, or event history analysis, we are interested in time between two events. Examples are time from diagnosis until death (or some other event), duration of hospitalization, or duration of unemployment. Formally, we are interested in a non-negative

* Corresponding author. Tel.: +386 1 543 7785; fax: +386 1 543 7771.
E-mail address: maja.pohar@mf.uni-lj.si (M. Pohar).

0010-4825/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2007.04.010

random variable T. The problem in practice is that T is often right censored, meaning that the final event is not observed at the time of analysis. Then what we really observe is $T^* = \min(T, C)$ where C is a censoring variable. This causes problems and requires special methods for analysis. Although we define here some of the basic quantities used in survival analysis we assume in what follows that the reader is familiar with at least the basic methods of survival analysis.

The main goal of survival analysis is to estimate the *survival function* $S(t) = P(T > t) = 1 - F(t)$ where $F(t)$ is the cumulative distribution function of T. Its graph is called the *survival curve*. Often we are interested in the conditional survival function $S(t|X)$ where X is a vector of covariates influencing the survival. This is usually done using one of the regression approaches, the Cox model [1] being the most common.

Most of the regression approaches model the *hazard function*, defined as

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}.$$

The motive for this is that the hazard is defined via conditional probability $P(t \leqslant T < t + \Delta t | T \geqslant t)$ which can be estimated even with censored data, and that knowing the hazard function also means knowing the survival function, since we have

$$S(t) = \mathrm{e}^{-\int_0^t \lambda(u)\,\mathrm{d}u}. \tag{1}$$

The Cox model, for example, assumes that

$$\lambda(t, z) = \lambda_0(t)\mathrm{e}^{\beta' z}, \tag{2}$$

where z is a vector of covariate values of an individual, and β is a vector of coefficients.

To motivate the relative survival approach assume now that the sample we want to analyse contains data about the survival of women with breast cancer from all parts of Slovenia (Slovenia being as good an example as any other country). And assume that our analysis has shown that women from the east do worse than women from the west. We might then be tempted to search for particular prognostic factors that could explain the difference. But, and this is really important, women from the east have lower life expectancy than women from the west, the difference being around 4 years. Thus, it may well be that our findings simply reflect this fact, and that *relatively* there is no difference in survival of women with breast cancer with respect to the geographical location. And even if there is, it will be smaller than our original analysis suggests. What we would then do is to calculate the observed and *expected* survival of women from the east and the west separately, and compare respective observed and expected curves. Of course, we can only calculate the expected survival if we have relevant population tables. Many countries, and we talk about this later, have population mortality tables broken down by calendar year, gender, and age. Some countries, like USA, will also use race, some will have regional information. National statistical offices will often be able to produce more detailed tables if needed, but in practice we are usually satisfied with gender, age, and year.

3. Relative survival

The cumulative relative survival function is defined [2] as

$$r(t) = \frac{S_{\mathrm{O}}(t)}{S_{\mathrm{P}}(t)}, \tag{3}$$

where $S_{\mathrm{O}}(t)$ denotes observed survival and $S_{\mathrm{P}}(t)$ stands for population or expected survival, which is estimated on the basis of population mortality tables. Obviously, $r(t)$ can be any non-negative number, although the methods are most often applied to data where $r(t)$ is less than 1. Correct calculation of the expected survival is not a straightforward task, and it is now generally accepted that the method of Hakulinen [3], gives the best results. Library `survival` in S-plus and R [4,5] has the method implemented (function `survexp`).

Relative survival function $r(t)$ will of course also depend on covariates, and there are different ways of modelling such dependence. We briefly review the main approaches.

3.1. Additive hazard models

Under the additive model the observed hazard (O) is a sum of the population hazard (P) and a non-negative excess (E) term

$$\lambda_{\mathrm{O}}(t, z) = \lambda_{\mathrm{P}}(t, c) + \lambda_{\mathrm{E}}(t, z), \tag{4}$$

where $z = (c, x)$, c denotes the vector of values of variables by which population tables are stratified, and x is a vector of values of any additional covariates we might want to include in the regression analysis. The excess hazard is usually modelled as

$\lambda_E(t,z) = h_0(t)\exp(\beta z)$ so that (4) becomes

$$\lambda_O(t,z) = \lambda_P(t,c) + h_0(t)e^{\beta z}. \quad (5)$$

Here $h_0(t)$ stands for baseline excess hazard. Eq. (4) gives a multiplicative relationship for the survival functions

$$S_O(t,z) = S_P(t,c)r(t,z),$$

where $r(t,z) = \exp\{-\int_0^t \lambda_E(u,z)\,du\}$. This equation is in the same form as (3), hence the phrase "relative survival model" is sometimes used specifically for the additive hazard class. Note that (4) assumes $\lambda_O(t,z) \geqslant \lambda_P(t,c)$ at all times and for any values of covariates, and $r(t,z)$ is a proper survival function. This will often be true, especially in cancer research. On the other hand, there can be subgroups of subjects that do better than the population, in which case the model would not be a good choice.

3.2. Multiplicative models

A general multiplicative hazard model is

$$\lambda_O(t,z) = \lambda_P(t,c)v(t,z). \quad (6)$$

The unit-free factor $v(t,z)$ can be seen as relative mortality and for this reason models of this type are sometimes called relative mortality models. The model has fewer mathematical restrictions than the additive model (4) but there have been arguments that additive models can be more realistic in practice, for cancer studies at least [6,7]. How good model (6) will fit depends of course on what we assume for $v(t,z)$, but the most commonly used form [8] is

$$\lambda_O(t,z) = \lambda_P(t,c)v_0(t)e^{\beta z}, \quad (7)$$

where $v_0(t)$ is an unspecified baseline relative mortality. Fitting is straightforward by including population mortality rates as a time-dependent covariate in the Cox model.

3.3. Individual relative survival

Relative survival, as described above, is about the group experience. The methods do not answer a very natural question "How long, relative to the general population, has a certain person lived?" Or "Did A live relatively longer than B"? Take for example Frank Sinatra who died in 1996 at the age of 81 years. French mathematician, Adrien Marie Legendre, died at the same age, but in the year 1833. Did one live relatively longer than the other? This question can be answered by calculating the average age of men who died in the same year in the same country if such data are available. For Sinatra and Legendre we find from the tables, described in Section 5, that respective averages were 73.2 and 38!! We can be satisfied with that, but we might also want to have a more sensitive measure, a measure that would tell us by how much one has outlived the other. To give another example, take Elvis Presley, who died in 1977 at the age of 42.6, and French writer, Guy de Maupassant, who was almost exactly the same age (42.9) at the time of his death in 1893. Elvis lived much less than the average which was 69.4 at the time in US, but Maupassant lived just slightly over the average in France in that year which was 42.03 years. Obviously, looking at the relative survival, Maupassant outlived Elvis by a lot, but can we say more?

We introduced such a measure in a recent paper [9]. The idea is to calculate for every individual the expected proportion of the general population that would not survive his/her measured time. Formally, let F_c denote the background population residual lifetime cdf for a given age, sex, and cohort year. Then the quantity we are interested in is the transformation $y = F_c(t)$ which converts the actual survival time t to the associated value y on the cdf. It then makes sense to take as an outcome measure the random variable $Y = F_c(T)$, which measures how long, after the event, and relative to the respective population, a person has lived. If T is censored, so is Y. For any given t, age, sex, and cohort year, the respective y is calculated from population tables. Fig. 1 shows calculation of y for Elvis Presley and Maupassant. The two curves represent the population cumulative distribution functions for the Americans born in 1935 (the year of Elvis's birth) and French born in 1850 (French tables go even further back!). We read from them that only 14% of the Americans had died until the age of Elvis' death, while Maupassant outlived 50% of the French in his time. In other words, his age was exactly the median in those times. Of course, the main reason for this big difference is the high infant mortality in France in that period, reflected in the steep rise of the cumulative distribution function in the first few years. If we condition our calculations on having survived for 5 years, the difference is smaller (Fig. 2), and we see that only 31% of the respective population died before Maupassant. Elvis outlived only 7% of those who survived the first 5 years of their lives. We explain in Section 7 how to create Figs. 1 and 2.

The biggest advantage of such an approach is that we get rid of the population differences without assuming any relation between λ_O and λ_P. Any regression model for Y can of course be used, but the choice will most often fall on the Cox model.

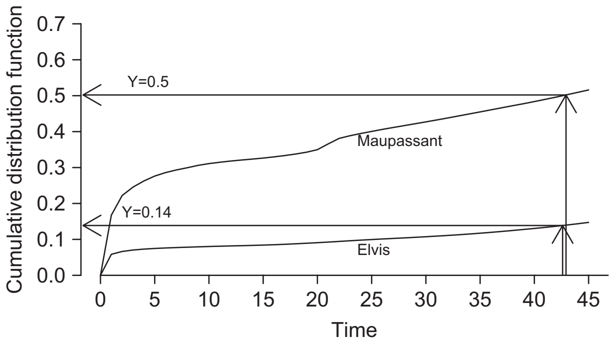


Fig. 1. Cumulative distribution functions for lifetime of men born in France in 1850 (Maupassant) and in USA in 1935 (Elvis). Arrows illustrate the computation of the outcome measure.

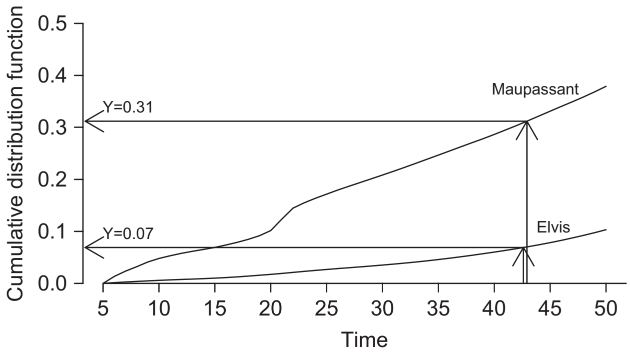


Fig. 2. Cumulative distribution functions for residual lifetime of men who were 5 years old in 1855 in France (Maupassant) and in USA in 1940 (Elvis). Arrows illustrate the computation of the outcome measure.

4. Estimation

Fitting any model to the transformed data presents no new problems, and as mentioned fitting the multiplicative model is also straightforward. Additive model on the other hand has its own peculiarities. The problem is that the existing methods require baseline excess hazard in (5) to be specified. This is usually done by assuming the baseline excess hazard to be constant in prespecified intervals. Noncritical usage of such an assumption can lead to problems. First, if the assumption is wrong, the fit of the model will not be good, and it will be impossible to show that it is the baseline hazard causing the bad fit. Second, estimates of the coefficients could be affected but we will not know about it. Third, it is of course desirable to have a correct estimate of the baseline excess hazard, even if the first two problems did not exist. Introduction of splines by Giorgi et al. [10] and fractional polynomials by [11] are useful improvements in this respect, but Pohar [12] has recently suggested even more flexible nonparametric approach that we outline here.

We first note that if we have cause specific information we consider deaths that are due to population hazard as censored and model (5) becomes the usual Cox model

$$\lambda_{\mathrm{O}}(t, z) = h_0(t)\mathrm{e}^{\beta' z}.$$

In the Cox model $h_0(t)$ is not needed for estimation of the coefficients, but it can be nonparametrically estimated.

On the other hand, if we know the parameters of the model, we can estimate the probability of disease specific death by $\lambda_{\mathrm{E}}/\lambda_{\mathrm{O}}$.

The idea for the fitting algorithm is thus to specify some initial values for the disease specific censoring indicator and then iterate between fitting a cause specific Cox model and estimating the probability of disease related death.

This procedure yields results that are free of any baseline excess hazard assumptions and can additionally provide information about the true form of the baseline excess hazard form.

5. Population tables available on the internet

Population tables are an indispensable part of any relative survival analysis. They can be obtained from the national statistical offices, but usually come in quite different formats depending on the national mortality and census data organization. Some work might then be needed for transforming such tables into a format required by R programs. Fortunately, web sites now exist that provide tables for various countries in a uniform format. One such site is the human mortality database (HMD, http://www.mortality.org). There are 26 countries included in this project at the moment and the tables needed for relative survival can be simply read as a table into any statistical package. For us, the column named "$q(x)$" (probability of death between exact ages x and $x+1$) will be of the main interest. However, the requirements of the HMD are such that only the countries where death registration and census data are virtually complete are included.

Another project, human lifetable database (HLD: http://www.lifetable.de/) is an even larger collection constructed by individuals or institutions using a variety of techniques. There are tables for 38 countries, and the time spans are in most cases longer than in HMD (in France going back to year 1806). The tables, however, are not directly comparable as they are given in a variety of formats and were calculated using different techniques. Again, most tables are split by sex, age, and calendar time and can be downloaded in the .txt format.

The described internet databases thus provide us with easily accessible population tables. However, the calculation of $\lambda_P(t,c)$ or $F_c(t)$ is still rather cumbersome, as the values must be calculated separately for each individual and each calendar year. The R package `relsurv`, described in the next section, greatly simplifies these procedures and enables automatic calculation from population tables in any format.

6. The R package `relsurv`

In previous sections, various relative survival methods and their accompanying problems have been presented. The `relsurv` package, that can be downloaded from CRAN [5], provides a software tool that simplifies their use. While the core functions are described in [13], the package has now been enriched with new methods and several functions that make it more user friendly. Two important additions, the EM-based estimation and the usage of the population tables, available on the Internet, are discussed in this paper. In this section we briefly review the basic functions, and explain how these additions are incorporated into the package.

There are two main advantages of the R `relsurv` package over other existing functions in various statistical softwares:

(i) all the main relative survival regression models (see Section 3) can be used, and the usage conforms to R standards;
(ii) any format of the population tables can be straightforwardly used in the functions.

6.1. *Regression model functions*

The core of the package are three functions that fit the models described in Section 3:

- `rsadd` fits the additive model (4).
- `rsmul` fits the multiplicative model (6).
- `rstrans` fits the Cox model for the transformed variable Y (Section 3.3). If only the transformation times are needed, this can be done directly by the `survexp` function (`survival` package) or by function `rstrans`, where the transformed times are returned in output value `y` (`fit <- rstrans(...), y <- fit$y`).

All the functions follow the same syntactic rules, here is an example of a basic call to function `rtrans`:

```
rstrans(formula, data, ratetable)
```

The left-hand side of the `formula` must be a `Surv` object. For example, if `time` and `status` are the survival time variable and the censoring indicator, and x is a covariate, then the command may be

```
rstrans(Surv(time,status)~x)
```

Apart from the observed data set, which we will pass to the function as argument `data`, the population mortality table has to be specified in the argument `ratetable`. The population mortality tables have to be organized as an object of class `ratetable` (defined in package `survival`), default is the `survexp.us` table that contains the US data (also in the `survival` package). The usage of population tables is further described in Section 6.2.

Each model also allows for some specific arguments, in particular, the additive model function `rsadd` additionally has options for different estimating methods, an example of a call with the new estimating method, outlined in Section 4, is

```
rsadd(Surv(time,status)~x,method='EM')
```

Explanation for other parameters in the function calls can be found in [13], and in the help file. The default values will do in most cases.

6.2. *The population tables*

To simplify the usage of the population tables in R, a special object named `ratetable` has been introduced in package `survival` [14]. A table containing conditional survival probabilities is given attributes that describe its organization and thus enable other functions to use it in the calculations. The main function for dealing with the `ratetable` in R is called `survexp` and this function is also used in all the `relsurv` package calculations.

Once the tables are organized as the `ratetable` object, they can be directly used in all the relative survival functions without having to care about their organization (i.e. the variables they are broken by). Instructions on how to put a table into the proper format are given by [14] and summarized in [13], but to make the methods as straightforward as possible we here introduce the R functions (included in `relsurv` package) that simplify this process even further:

- `transrate.hmd`: this function reads the tables in .txt format obtained from the HMD site (see Section 5) and puts them into the `ratetable` object.
  There are different tables available for each country, those that we need are called "Life tables", and we need them by year of death (period). The tables are given separately for males and females and our program works with those broken by yearly intervals with respect to both age and calendar year (1 × 1). The tables should be saved in .txt format with the first line (title) deleted. Then we can, for example, write:

  ```
  poptab <- transrate.hmd(male = "mltper_1x1.txt", female = "fltper_1x1.txt")
  ```

  The uniformity of the methods used in the HMD makes it certain for us to have been given population tables broken by age (in yearly intervals from 0 to 110), sex and calendar time (in yearly intervals as well).
  The outcoming object `poptab` is of the class `ratetable` (this can be checked by writing `is.ratetable(poptab)` and its organization can be summarized by `summary(poptab)` or inspected in more detail by writing `attributes(poptab)`.
- `transrate.hld`: this function transforms the HLD tables. These tables can again be downloaded in .txt format with both sexes included in each file. Typically, however, each calendar time period is stored in a separate table, the main goal of this function being to join these tables and see to the different formats. The Finnish tables from 1975 to 1995 for example can be formed by

  ```
  finpop <- transrate.hld(c("FIN_1971-75.txt","FIN_1976-80.txt",
        "FIN_1981-85.txt","FIN_1986-90.txt","FIN_1991-95.txt"))
  ```

  There are two optional arguments provided for this function. The first is named `cut.year` and allows the user to specify the cut points for calendar time when the years covered by the tables aree not consecutive. The other is named `race` and is of the same length as the argument `file`. It allows the user to create a fourth dimension of the ratetable (here named race) by specifying which of the files belongs to which race. This parameter could of course also be used to join the tables belonging to two different countries, but under the condition that their organization as well as the lapse of time they cover are identical.
  An example of the usage of these two parameters are the New Zealand tables which are given separately for the Maori and non-Maori population (the argument `cutyear = c(1980,1985)` specifies that the 1980–1982 tables should be used up to 1985).

  ```
  nzpop <- transrate.hld(c("NZL_1980-82_Non-maori.txt",
        "NZL_1985-87_Non-maori.txt","NZL_1980-82_Maori.txt",
        "NZL_1985-87_Maori.txt"),cut.year = c(1980,1985),
        race = rep(c("nonmaori","maori"),each = 2))
  ```

- `joinrate`: a function that helps constructing a joint `ratetable` object from two or more tables that are broken by age, sex and year by adding a new dimension. Such a table could readily be used in a model where, for example, individuals from different countries are to be compared. If the comparison is to be proper, only the common cut points can be used. Warnings are issued if the originals are not organized in a comparable way.
- `transrate`: a simple function that could be of assistance while forming a `ratetable` object from two R tables (for males and females).

Important: Note that the hazards in a `ratetable` object are expressed in units 1/day. Therefore, all the times and ages used in any `relsurv` function must be expressed in days and the dates must be in the `date` format.

7. Examples

In this section we present three examples of the usage of the package, with emphasis on the usage of population tables. We first explain how to create figures from Section 3.3, and then present two examples of relative survival analysis.

7.1. Constructing Figs. 1 and 2

We start by describing how the data for plotting Figs. 1 and 2 were obtained. The population tables for the two countries in question that include the required cohort years must be downloaded from the human life table database and saved in .txt files.

The function `transrate.hld` transforms them into the R format:

```
frpop <- transrate.hld("FRA18061997.txt")
uspop <- transrate.hld("USA19011999.txt")
```

To calculate the cumulative distribution function for the first 45 years after Elvis's birth we write

```
y <-1-survexp(~ratetable(age = 0,sex = "male",year = as.date("8Jan1935")),
times = (0:45)*365.24,ratetable = uspop)$surv
```

and then simply plot y versus age

```
plot(0:45,y,type = "l")
```

7.2. Survival after myocardial infarction

In this section we present a study of patients' survival following an acute myocardial infarction (AMI). The data were collected at the University Clinical Center in Ljubljana and refer to 1040 patients diagnosed between 1982 and 1986 and followed up until the year 1997. During this time 547 deaths occurred and as the causes of death are unknown, this is a good example of the need of the relative survival methodology. We will use the Cox model on transformed times (Section 3.3).

We have two data sets—the first, included in the `relsurv` package under the name `rdata`, contains the observed survival times, censoring status and covariate values for our patients. The second, a `ratetable` object `slopop` (also included in the `relsurv` package), contains the population mortality tables for Slovenia. It is broken down by age, sex, and cohort year and contains yearly intervals. To load the two data sets into our R workspace, we write

```
> library(relsurv)
> data(slopop)
> data(rdata)
```

The goal of the study is to evaluate the impact of sex, age, and calendar year on the survival after AMI. A glimpse of the data tells us how the variables are organized and named

```
> rdata[1:2,]
       time    cens    age    sex    year       agegr
1      2657    1       68     2      24Jun82    61-70
2      1097    1       63     2      31Aug82    61-70
```

We can see that the survival time and the status are stored under the names "time"and "cens", respectively, the survival time being already quoted in days as required by all the relative survival functions. The "age" variable, however, is measured in years and we have to change it to days (see remark at the end of Section 6.2). The easiest way to do this is to specify the change in the `ratetable` part of the formula:

```
rstrans(Surv(time,cens)~age+sex+year+ratetable(age = age*365.24,
  sex = sex,year = year), data = rdata, ratetable = slopop)
```

The obtained results are given in Table 1.

We can see that all the three variables significantly influence relative survival. Survival is better for men than women and it improves with calendar year implying that the health system has improved for AMI patients in this period. It is not surprising that age is negatively associated with the outcome in such analysis as older patients will be losing relatively less than the younger ones.

7.3. Survival of patients with colon cancer

The other example we discuss is a colon cancer study based on the data from the Finnish Cancer Registry. The data set consists of 6247 patients diagnosed in the period of 1975–1994 with follow-up to the end of 1995, and can be downloaded from the

Table 1
The results of the model for AMI data

	β	$\exp(\beta)$	SE	z	p
Age	−0.0139	0.986	0.0049	−2.83	0.005
Sex	0.5287	1.697	0.1010	5.24	< 0.001
Year	−0.0002	1.000	0.0001	−2.52	0.012

Table 2
The results of the model for colon carcinoma data

	β	$\exp(\beta)$	SE	z	p
Year [85,94]	−0.313	0.731	0.075	−4.192	<0.001
Sex	−0.048	0.953	0.077	−0.623	0.533
Agegrp [45,59]	−0.142	0.867	0.156	−0.913	0.361
Agegrp [60,74]	1.058	0.057	0.143	0.397	0.691
Agegrp [75+]	0.299	1.349	0.151	1.989	0.047
fu[0, 1]	−2.514	0.081	0.148	−16.933	<0.001
fu(1, 2]	−2.705	0.067	0.153	−17.707	<0.001
fu(2, 3]	−2.908	0.055	0.160	−18.172	<0.001
fu(3, 4]	−3.135	0.044	0.173	−18.080	<0.001
fu(4, 5]	−3.292	0.037	0.188	−17.474	<0.001

internet site http://www.pauldickman.com. We make use of the `relsurv` package to reproduce the analysis done in [6] where the effect of three covariates, i.e. age, sex, and year in the first 5 years was studied. Their analysis was performed on grouped data and we will keep the same model in order to get comparable results. Again we inspect the organization of the data:

```
> colon[1:2,]
sex      age   status   year8594         agegrp   time   diag
Female   78    1        Diagnosed 75-84  75+      2511   7Oct78
Male     80    1        Diagnosed 75-84  75+      259    7Apr80
```

To get the population tables we proceed as described in Section 4—we download the required cohort years (1975–1995) from the HLD web site and use the `transrate.hld` function (the command is explicitly given in Section 6.2). By checking the attributes of this rate table (`attributes(finpop)$dimnames`), we see that sex is defined as a factor variable with levels “male” and “female”, and the easiest way to proceed in our case is therefore to use an integer with values 1 and 2 for males and females, respectively. The rest is done the same as in the first example:

```
fit.col <- rsadd(Surv(time,status)~year8594+sex+agegrp+
     ratetable(age=age*365.24,sex=as.numeric(sex),year=diag),
     data=colon, ratetable=finpop, int=5)
```

The output is given in Table 2 and is identical to the third column of colon carcinoma results in [6, Table 1]. The other columns in that table can be reproduced by changing the argument `method` of the `rsadd` function to the desired option.

We could plot the follow-up intervals coefficients given in Table 2 to get an idea of the baseline excess hazard behaviour. A much more flexible and reliable estimate can be obtained by plotting the results after fitting using the EM method

```
fit.col2 <- rsadd(Surv(time,status)~year8594+sex+agegrp+
     ratetable(age=age*365.24,sex=as.numeric(sex),
     year=diag),data=colon,ratetable=finpop,int=5,method="EM")

plot(lowess(fit.col2$times,fit.col2$lambda0,f=0.1),type="l")
```

The results are presented in Fig. 3. We can see that the baseline excess hazard is very high in the first half year after the diagnosis, implying that this period is the most critical for dying due to cancer. The excess hazard then decreases, but remains non-negligible compared to the population hazard throughout the 5-year time period, so a constant risk of dying due to colon cancer remains present.

8. Discussion

Relative survival analysis offers answers to questions that cannot be answered with standard analysis. For now, the methods are used extensively in cancer registries, but are almost unknown to other medical professionals. For example, no textbook on survival analysis has a chapter, or even a section, devoted to the topic. This is partly due to the fact that relative survival analysis is inseparably tied to the population data, which is seen as an obstacle to its usage. With ever greater availability of population mortality tables, this is bound to change, and existence of relevant software is of the key importance for doing so. Even more so if it comes as an “all inclusive” package, capable of performing all (or most at least) the analysis, and handling of the data. The package `relsurv`, presented in this paper, provides all the necessary functions to perform a quality relative survival analysis. Additionally, it has an important capability of reading now widely available population tables. Its usage is simple, especially if one is already acquainted with the R language.

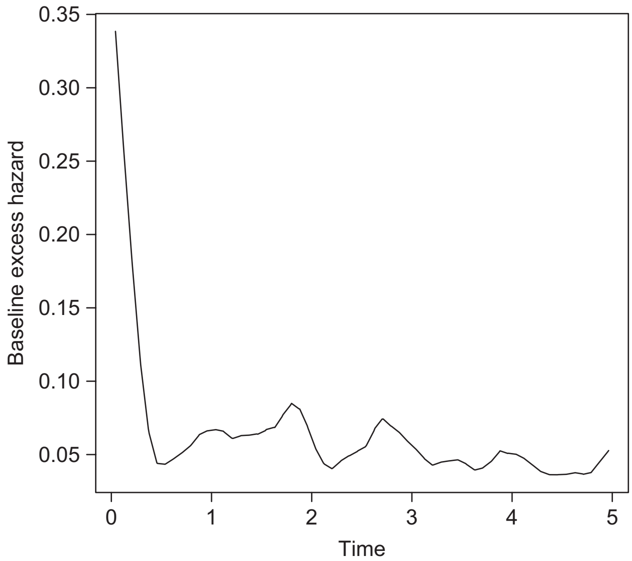


Fig. 3. The baseline excess hazard for the colon cancer data.

With new methods appearing, importance of relative survival methods is rising. Hopefully, this will lead to even more detailed population tables (regional data being one simple example), which in turn will give a new incentive to the field.

References

[1] D.R. Cox, Regression models and life-tables (with discussion), J. R. Statist. Soc. Ser. B 34 (1972) 187–220.
[2] F. Ederer, L.M. Axtell, S.J. Cutler, The Relative Survival Rate: A Statistical Methodology, vol. 6, National Cancer Institute Monograph, 1961, pp. 101–121.
[3] T. Hakulinen, L. Tenkanen, Regression analysis of relative survival rates, J. R. Stat. Soc. Ser. C 36 (1987) 309–317.
[4] Mathsoft, S-Plus 2000 Guide to Statistics 2, Data Analysis Products Division of MathSoft, Seattle, WA, 1999.
[5] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2005. URL ⟨http://www.R-project.org⟩. ISBN 3-900051-07-0.
[6] P.W. Dickman, A. Sloggett, M. Hills, T. Hakulinen, Regression models for relative survival, Stat. Med. 23 (2003) 51–64.
[7] J.D. Buckley, Additive and multiplicative models for relative survival rates, Biometrics 40 (1984) 51–62.
[8] P.K. Andersen, K. Borch-Johnsen, T. Deckert, A. Green, P. Hougaard, N. Keiding, S. Kreiner, A Cox regression model for relative mortality and its application to diabetes mellitus survival data, Biometrics 41 (1985) 921–932.
[9] J. Stare, R. Henderson, M. Pohar, An individual measure of relative survival, J. R. Stat. Soc. Ser. C 54 (2005) 115–126.
[10] R. Giorgi, M. Abrahamowicz, C. Quantin, P. Bolard, J. Estève, J. Gouvernet, J. Faivre, A relative survival regression model using b-spline functions to model non-proportional hazards, Stat. Med. 22 (2003) 2767–2784.
[11] P.C. Lambert, L.K. Smith, R.J. Jones, J.L. Botha, Additive and multiplicative covariate regression models for relative survival incorporating fractional polynomials for time-dependent effects, Stat. Med. 24 (2005) 3871–3885.
[12] M. Pohar, J. Stare, EM algorithm based estimation in relative survival regression, in: L. Edler, D. Warne (Eds.), Annual Conference of the International Society for Clinical Biostatistics, 27–31 August 2006, Geneva, Abstract book, International Society of Clinical Biostatistics, Geneva, 2006, p. 85.
[13] M. Pohar, J. Stare, Relative survival analysis in R, Comput. Methods Programs Biomed. 81 (2006) 272–278.
[14] T. Therneau, J. Offord, Expected survival based on hazard rates (update). Technical Report 63, Section of Biostatistics, Mayo Clinic, 1999.

Maja Pohar has graduated from mathematics at the University of Ljubljana and is now a final year Ph.D. student of statistics at the same university. She is the author of six SCI papers and has written papers on relative survival in the Journal of the Royal Statistical Society—Series C, Statistics in Medicine, and Computer Methods and Programs in Biomedicine. She is the author of the R package relsurv that is included in CRAN and provides functions for regression in relative survival.

Janez Stare is a full profesor at the Faculty of Medicine at the University of Ljubjana and the head of the Department of Biomedical Informatics.
He is the author of more than 25 SCI papers and has recently written papers on relative survival in the Journal of the Royal Statistical Society—Series C, Statistics in Medicine, and Computer Methods and Programs in Biomedicine.
He has been an invited speaker on the topic of relative survival at the Annual Conference of the International Society for Clinical Biostatistics in Leiden, 2004, and at the ROeS Seminar in Graz, 2005.