Biometrics 000, 1–19 DOI: 000
000 0000

ON ESTIMATION IN RELATIVE SURVIVAL

Maja Pohar Perme[1,*], Janez Stare[1] and Jacques Estève[2]

1. Department of Biostatistics and Medical Informatics, University of Ljubljana, Vrazov trg 2, SI-1000 Ljubljana,Slovenia

2. Université Claude Bernard, Hospices Civils de Lyon, Service de Biostatistique, 162, Ave Lacassagne 69424 Lyon Cedex 03, France

∗ email: maja.pohar@mf.uni-lj.si

Summary: Estimation of relative survival has become the first and the most basic step when reporting cancer survival statistics. Standard estimators are in routine use by all cancer registries. However, it has been recently noted that these estimators do not provide information on cancer mortality that is independent of the national general population mortality. Thus they are not suitable for comparison between countries. Furthermore, the commonly used interpretation of the relative survival curve is vague and misleading. The present paper attempts to remedy these basic problems. The population quantities of the traditional estimators are carefully described and their interpretation discussed. We then propose a new estimator of net survival probability that enables the desired comparability between countries. The new estimator requires no modelling and is accompanied with a straightforward variance estimate. The methods are described on real as well as simulated data.

Key words: Age-standardization; Cancer registry data; Competing risks; Net survival; Relative survival; Survival analysis

1. Introduction

Survival probability of cancer patients has been used for many years as one of the main tools for evaluation of therapeutic advances. With improved treatments and prognosis, studies often now have long follow-up times and it is common to have a substantial proportion of deaths from causes other than the cancer under study. In the usual situation, the cause of death is unavailable or unreliable. Hence the field of relative survival has developed in which observed deaths are compared with those expected from general population life-tables.

We distinguish two goals:

(1) To compare the observed survival (S_O) to the survival of a disease-free group having the same demographic characteristics as the study group. If the expected general population survival is S_P, the comparison is made using the ratio

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

This has been named the *relative survival ratio*.

(2) To estimate survival in the hypothetical situation where the disease under study would be the only possible cause of death.

This estimation is made possible by decomposing the observed hazard into the hazard due to the disease and that due to other causes. This decomposition can be carried out if the time to death due to the disease and the time to death due to other causes are conditionally independent given a known set of covariates. We then assume that the hazard due to other causes is given by the population mortality and therefore, that the observed hazard is larger than the population hazard. Under this assumption, we use the term excess hazard for the hazard due to the disease and we have the relation

$$\textit{observed hazard} = \textit{population hazard} + \textit{excess hazard}. \tag{1}$$

A survival function derived from the excess hazard alone is termed the *net survival*.

In this paper, we shall focus on three estimators in widespread use in relative survival. We will refer to them as Ederer I (Ederer et al., 1961), Hakulinen (Hakulinen, 1982) and Ederer II (Ederer et al., 1961, page 110) and define them in Section 3 below.

While all three were proposed as variants for estimating the denominator of the relative survival ratio, they have also been interpreted as estimators of net survival. We will show that this is not generally correct. When the excess and the population hazard are not affected by any common covariates, all methods to be discussed in this paper estimate the same population quantity. In practice however, excess hazard almost always depends on demographic variables (e.g. age) and one of the aims of this paper is to find the population quantities that the estimators are estimating in such situations (Sections 2 and 3).

While the concept of net survival may seem too hypothetical to be of interest by itself, it becomes crucial when trying to compare cancer burden between countries, since it is independent of the general population mortality. Up to now, most authors who have produced large sets of survival statistics (Sant et al. (2009) and references therein) have calculated age-standardized relative survival using the Hakulinen method, a variant of the Ederer method and an ad-hoc method of standardization. It has been noted recently that this approach produces biases and inconsistencies, but a good method for correction of these deficiencies has not yet been proposed (Pokhrel and Hakulinen, 2009). An obvious solution to the problem of comparability is to use multivariate models that enable estimation of excess hazard in homogeneous groups and then average the obtained survival curves. Common ways to model the excess hazard include assuming proportional hazards with either a parametric (Hakulinen and Tenkanen, 1987; Estève et al., 1990; Dickman et al., 2004; Nelson et al., 2007) or non-parametric (Pohar Perme et al., 2009; Sasieni, 1996) baseline excess hazard, or adopting a fully additive model following Aalen (Aalen et al., 2008; Cortese and Scheike, 2008; Zahl

and Aalen, 1998). Any approach that includes modelling is of course dependent upon the validity of assumptions made in the model. The main aim of this paper is to introduce a new estimator of net survival that does not require modelling. We do this in Sections 4 and 5.

2. The population values

We shall study the population quantities of interest in relative survival. First, we introduce some notation.

Assuming the additive model (1) for each individual i in the diseased cohort, we define T_{Ei} to be the time to death due to the disease of interest, T_{Pi} to be the time to death due to other causes operating in the general population, and C_i to be the time to censoring. Since death from one cause precludes observing the time to death due to other causes, we cannot observe both T_{Ei} and T_{Pi}, but rather observe only $T_i = min\{T_{Ei}, T_{Pi}\}$ (subject to censoring by C_i). Let $U_i = min\{T_i, Ci\}$ denote the follow-up time on individual i, and define failure indicator δ_i equal to 0 if the death is censored ($T_i > C_i$) and 1 otherwise. Further, let X_i denote covariates and let D_i denote a subset which we will describe as demographic variables (usually age, sex and year). The observed data on individual i are then (U_i, δ_i, X_i). We define the survival functions $S_{Ei}(t) = P(T_{Ei} > t|X_i)$ and $S_{Pi}(t) = P(T_{Pi} > t|D_i)$ and let $\lambda_{Ei}(t)$ and $\lambda_{Pi}(t)$ denote the corresponding hazard functions. We assume T_E and T_P to be conditionally independent given D. Further, we assume non-informative censoring, i.e. $S_C = S_{Ci} = P(C_i > t)$ is common for all i. We use the symbol Λ (subscripted as appropriate) to denote a cumulative hazard function. For simplicity, we shall assume a finite population of diseased patients of size N. The size of a sample from this population shall be denoted by n.

In the *relative survival setting* we have two sets of data: data on time to death (from all causes) in a cohort of patients with the disease of interest ($\{U_i, \delta_i, X_i\}$, for $i = 1, \ldots n$); and population mortality tables that provide the population all-cause hazard functions for various

demographic strata in the general population. We assume that within each demographic stratum defined by the values of D, the set of patients in the diseased cohort is comparable to the general population with respect to risk factors related to the causes of death operating in the general population, and we use the tables to get $\lambda_{Pi}(t)$ values for each individual i at each time t. We further assume that the disease is "rare", i.e. that removing disease of interest from population life tables would have negligible effect on λ_{Pi}. To shed more light on some of the concepts, we shall also mention the *cause specific setting*, where the cause of death is known.

The *net survival* is the survival associated with the *excess (net) hazard* working alone

$$\lambda_E(t) = \lim_{\Delta t \to 0} \frac{P(t < T_E \leq t + \Delta t | T_E > t)}{\Delta t}. \tag{2}$$

Using the relationship between the hazard and survival function, the net survival function for an individual is defined as $S_{Ei}(t) = \exp\{-\int_0^t \lambda_{Ei}(u)du\}$ and the overall (marginal) net survival equals $S_E(t) = (1/N)\sum_{i=1}^N S_{Ei}(t)$. The associated hazard λ_E that corresponds to this quantity, i.e.

$$S_E(t) = \exp\{-\int_0^t \lambda_E(u)du\}, \tag{3}$$

is given by

$$\lambda_E(t) = \frac{\sum_{i=1}^N S_{Ei}(t)\lambda_{Ei}(t)}{\sum_{i=1}^N S_{Ei}(t)}. \tag{4}$$

The hazard for the overall net survival (3) is thus a weighted average of the individual hazards, with weights equal to the probability that an individual is still alive in the hypothetical world of the disease being the only cause of death.

In practice, the net survival is not observable when there are competing causes of death.

We define

$$\lambda_E^*(t) = \lim_{\Delta t \to 0} \frac{P(t < T_E \leq t + \Delta t | T > t)}{\Delta t}, \tag{5}$$

and stress that the conditioning in (5) depends on $T = \min\{T_E, T_P\}$ rather than T_E alone as in (2). This λ_E^* is the *cause specific hazard* (Tsiatis, 2005), when the population risk is also present. Similarly, $\lambda_P^*(t)$ is defined as

$$\lambda_P^*(t) = \lim_{\Delta t \to 0} \frac{P(t < T_P \leq t + \Delta t | T > t)}{\Delta t}$$

and the two quantities sum up into the observed hazard $\lambda_O(t) = \lambda_E^*(t) + \lambda_P^*(t)$. When considering the excess hazard λ_{Ei} of each individual and thus conditioning on D_i, the conditional independence of T_P and T_E given D implies that $\lambda_{Ei}^* = \lambda_{Ei}$. On the contrary, conditioning on $T_P > t$ becomes crucial when considering a cohort that differs with respect to D. Following the definition in (5), the cause specific hazard can be written as

$$\lambda_E^*(t) = \frac{\sum_{i=1}^N S_{Oi}(t)\lambda_{Ei}(t)}{\sum_{i=1}^N S_{Oi}(t)} = \lambda_O(t) - \frac{\sum_{i=1}^N S_{Oi}(t)\lambda_{Pi}(t)}{\sum_{i=1}^N S_{Oi}(t)}, \tag{6}$$

where the observed hazard is given by $\lambda_O(t) = \{\sum_{i=1}^N S_{Oi}(t)\lambda_{Oi}(t)\}/\{\sum_{i=1}^N S_{Oi}(t)\}$. As in (4), the hazard in (6) is calculated as a weighted average of the individual excess hazards, but the weights depend on the probability of surviving all causes and not just the excess risk. The observed survival of a cohort of patients is

$$S_O(t) = \frac{1}{N}\sum_{i=1}^N \exp[-\int_0^t \{\lambda_{Pi}(u) + \lambda_{Ei}(u)\}du] = \exp\{-\int_0^t \lambda_P^*(u)du\}\exp\{-\int_0^t \lambda_E^*(u)du\}.$$

and the term associated with the cause specific hazard λ_E^*

$$S_E^*(t) = \exp\{-\int_0^t \lambda_E^*(u)du\} \tag{7}$$

is in competing risk setting often referred to as the *observable net survival* (Marubini and Valsecchi, 1995), since it is directly estimable by the Kaplan-Meier method with deaths due to population risks regarded as censoring. However, since such censoring due to population risks is informative (common dependence on covariates D), the interpretation of the observable

net survival is unclear. The observable net survival $S_E^*(t)$ is not a proper survival function and cannot be written as the probability of some random variable exceeding a chosen time t. In particular, it cannot be used as a measure of cancer burden, since it depends on the general population mortality through S_{Oi} in (6).

The relative survival ratio

$$S_R(t) = \frac{\sum_{i=1}^{N} S_{Oi}(t)}{\sum_{i=1}^{N} S_{Pi}(t)} = \frac{\sum_{i=1}^{N} \exp\{-\int_0^t \lambda_{Oi}(u)du\}}{\sum_{i=1}^{N} \exp\{-\int_0^t \lambda_{Pi}(u)du\}} = \exp\{-\int_0^t \lambda_E^{**}(u)du\} \tag{8}$$

is also of interest. This involves a third quantity λ_E^{**} (not necessarily a hazard function), which is defined by

$$\lambda_E^{**}(t) = \lambda_O(t) - \frac{\sum_{i=1}^{N} S_{Pi}(t)\lambda_{Pi}(t)}{\sum_{i=1}^{N} S_{Pi}(t)}. \tag{9}$$

Note that λ_E as defined in (4) and λ_E^* as defined in (6) are necessarily non-negative, while this cannot be said of λ_E^{**} as defined in (9). A situation where λ_E^{**} is negative will result in increasing relative survival ratio despite the mortality in the study population being higher then in background.

Equations (4), (6) and (9) present three distinct hazard rates that lead to three distinct functions: the net survival (3), the observable net survival (7) and the relative survival ratio (8). If the excess hazard is the same for all individuals ($\lambda_{Ei} = \lambda_E$), the weights in (4), (6) and (9) cancel and all three equations result in the same population quantity λ_E. If not, then the net survival is the only one of these quantities that is independent of the population risks. In developing the early relative survival methodology, it seems that researchers were trying to estimate net survival, though as we shall show in the next section, they were instead estimating either the observable net survival or the relative survival ratio.

3. The existing estimators

In this section, we study the continuous time versions of the existing estimators in relative survival setting. Let $N_i(t) = I(T_i \leq t, T_i \leq C_i)$ and $Y_i(t) = I(T_i \geq t, C_i \geq t)$ denote the counting process and the at-risk process for each individual in the sample ($i = 1, \ldots, n$). Further, let $N(t) = \sum_i N_i(t)$ and $Y(t) = \sum_i Y_i(t)$. The intensity process for each individual following the additive model (1) can be written as $Y_i(t)\{\lambda_{Pi}(t) + \lambda_{Ei}(t)\}$. Assuming equal excess hazard λ_{Ei} for all i and following the same arguments as in the derivation of the Nelson-Aalen estimator (Aalen et al., 2008), the cumulative excess hazard can be estimated as (Andersen and Vaeth, 1989)

$$\widehat{\Lambda}^*_E(t) = \int_0^t \frac{dN(u)}{Y(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i(u) d\Lambda_{Pi}(u)}{Y(u)}. \tag{10}$$

Equation (10) is the continuous version of what has been named the *Ederer II* estimator (Ederer et al., 1961). The first term on the right-hand side equals the Nelson-Aalen estimator of the observed cumulative hazard Λ_0. The second term represents the cumulative population hazard $\widehat{\Lambda}^*_P$. For each interval between follow-up times, $d\widehat{\Lambda}^*_P$ is calculated as the average change in cumulative population hazard over that interval for patients who were still at risk through that interval. Assuming noninformative censoring ($S_{Ci} = S_C$) and noting that $E\{Y_i(t)\} = P\{Y_i(t) = 1\} = S_{Oi}(t)S_C(t)$, the term $(1/n)\sum_{i=1}^n Y_i(t)d\Lambda_{Pi}(t)$ converges to $(S_C(t)/N)\sum_{i=1}^N S_{Oi}(t)d\Lambda_{Pi}(t)$ and $(1/n)\sum_{i=1}^n Y_i(t)$ converges to $(S_C(t)/N)\sum_{i=1}^N S_{Oi}(t)$. The censoring term S_C cancels out and we can see that the Ederer II estimator is consistent for the cumulative hazard associated with (6) and estimates the observable net survival (7).

Using the same notation, the *Ederer I* estimator (Ederer et al., 1961) can be written as

$$\widehat{\Lambda}^{**}_E(t) = \int_0^t \frac{dN(u)}{Y(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i^{**}(u) d\Lambda_{Pi}(u)}{Y^{**}(u)}, \tag{11}$$

where $Y_i^{**}(t) = S_{Pi}(t)$. Since the first term estimates the cumulative observed hazard and the second term equals the cumulative version of the second term in (9), this estimator clearly estimates the cumulative hazard associated with relative survival ratio.

The construction of the *Hakulinen* estimator (Hakulinen, 1982) involves distinguishing between censoring due to the end of the initially planned follow-up and interim censoring. It is assumed that the initially planned follow-up time, referred to as the potential follow-up time and denoted by τ_i, is known for all individuals (e.g., time from individual i's entry into the study to study closeout). Individual i's actual censoring time is given by $C_i = \min(\widetilde{C}_i; \tau_i)$, where $\widetilde{C}_i$ is the time to interim censoring. To correct the estimator in the situation where the potential follow-up time and thus the censoring time is correlated with event time, the Hakulinen estimator extends the definition of Y^{**} in formula (11) to equal $Y_i^{**}(t) = S_{Pi}(t)I(C_i \geq t)$ if $\delta_i = 0$ and $Y_i^{**}(t) = S_{Pi}(t)I(\tau_i \geq t)$ if $\delta_i = 1$. In this way, the estimator attempts to diminish the bias caused by informative censoring in the first term on the right-hand of (11) by introducing a similar bias in the second term. If censoring C is independent of T, both methods estimate the same quantity, if there is no interim censoring and all potential follow-up times are all equal to (or larger than) the maximum observed time, the Ederer I and Hakulinen estimator coincide.

Using the relationship $S(t) = \exp\{-\int_0^t d\Lambda(u)\}$, both (10) and (11) can be written as a ratio of the observed survival and a term that is referred to as population survival. Such a ratio is a weighted sum of individual S_{Ei} (Estève et al., 1990) and can never estimate the net survival if these weights are not identical.

4. Estimation of net survival

We first explain the idea of the proposed net survival estimatior with the cause specific data and then adapt it for the relative survival setting.

Cause specific setting. In the cause specific setting, a solution to the problem of net survival estimation has been proposed by Robins (1993) and Satten et al. (2001).

The observed data in the cause specific setting can be represented by the counting process $N_{Ei}(t) = I\{T_i \leq t, T_i \leq C_i, T_{Ei} < T_{Pi}\}$ and the at risk process $Y(t)$. If both the population

and the excess hazard are affected by variables D, censoring by the population hazard is informative. The Nelson-Aalen estimator $\widehat{\Lambda}_E^*(t) = \int_0^t dN_E(u)/Y(u)$ is consistent for the cumulative version of (6) and gives a biased estimate of (4).

In order to get an unbiased estimate of $\Lambda_E(t)$ in the cause specific setting, we weight both the counting and the at risk process of each individual using his/her population survival time distribution. We define

$$N_{Ei}^w(t) = \frac{N_{Ei}(t)}{S_{Pi}(t_i-)},\ Y_i^w(t) = \frac{Y_i(t)}{S_{Pi}(t-)} \tag{12}$$

and propose the estimator $\widehat{\Lambda}_E^w(t) = \int_0^t dN_E^w(u)/Y^w(u)$, which is unbiased despite the censoring due to population hazard ($N_E^w = \sum N_{Ei}^w$ and $Y^w = \sum Y_i^w$, as usual).

Intuitively, the weighting of Y increases the sample size still at risk to account for the expected proportion of patients that may have been lost due to the population hazard. The weighting of N_E increases the number of events in the same manner. Since all the required information for the weights can be obtained from population tables, the weighting can be performed without modeling and no further assumptions regarding modeling and identifiability have to be made. Also, the t_i- in (12) can be replaced by t_i, as S_{Pi} is assumed continuous.

Relative survival setting. One option to estimate the net survival (3) in the relative survival setting is to use a multivariate relative survival model. Assuming the population risk is fully determined by the demographic covariates, the conditional independence assumption

$$P(t < T_E \leq t + \Delta t | T_P > t, T_E > t, D) = P(t < T_E \leq t + \Delta t | T_E > t, D)$$

implies that, assuming we can provide an adequate model for the excess hazard, the censoring due to population hazard is conditionally independent and net survival can be estimated as the average of the predicted values for each individual.

To estimate (4) directly, without using a model, we propose weighting the Ederer II estimator following the same logic as in (12). By letting $N_i^w(t) = N_i(t)/S_{Pi}(t)$ and $N^w(t) = \sum N_i^w(t)$,

we define the new estimator as

$$\widehat{\Lambda}_E(t) = \int_0^t \frac{dN^w(u)}{Y^w(u)} - \int_0^t \frac{\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)}. \tag{13}$$

To find the population value of the above estimator (when $S_{Ci} = S_C$ for all i), we first note that $E\{Y_i^w(t)\} = S_C(t)S_{Oi}(t)/S_{Pi}(t) = S_C(t)S_{Ei}(t)$ and $E\{dN_i^w(t)|\mathcal{F}_t\} = E\{dN_i(t)|\mathcal{F}_t\}/S_{Pi}(t) = Y_i(t)d\Lambda_{Oi}(t)/S_{Pi}(t)$, where $\mathcal{F}_t$ denotes the history, i.e. all available information up to time t. Since

$$\frac{1}{n}\sum_{i=1}^n \frac{Y_i(t)d\Lambda_{Oi}(t)}{S_{Pi}(t)} \to \frac{S_C(t)}{N}\sum_{i=1}^N \frac{S_{Oi}(t)d\Lambda_{Oi}(t)}{S_{Pi}(t)} = \frac{S_C(t)}{N}\sum_{i=1}^N S_{Ei}(t)d\Lambda_{Oi}(t),$$

the population value of $d\widehat{\Lambda}_E(t)$ in (13) equals (the terms S_C cancel out)

$$\frac{\sum_{i=1}^N S_{Ei}(t)d\Lambda_{Oi}(t)}{\sum_{i=1}^N S_{Ei}(t)} - \frac{\sum_{i=1}^N S_{Ei}(t)d\Lambda_{Pi}(t)}{\sum_{i=1}^N S_{Ei}(t)} = \frac{\sum_{i=1}^N S_{Ei}(t)d\Lambda_{Ei}(t)}{\sum_{i=1}^N S_{Ei}(t)} = d\Lambda_E(t).$$

We have thus shown that the newly proposed method provides a consistent estimator of (4). Note that unlike the traditional estimators, this method can not be expressed as a ratio of the observed and population survival.

5. Variance

The variance of the curves calculated by the Hakulinen and both Ederer methods is estimated by (Andersen and Vaeth, 1989)

$$\widehat{\sigma}^{*2}(t) = \int_0^t \frac{J(u)}{Y(u)^2} dN(u), \tag{14}$$

where $J(t) = I\{Y(t) > 0\}$ and $J(t)/Y(t)$ is interpreted as 0 whenever $Y(t)$ is 0.

Following similar arguments, the variance of the newly proposed method can be estimated by (see Appendix)

$$\widehat{\sigma}^2(t) = \int_0^t \frac{J(u)}{Y^w(u)^2}\sum_{i=1}^n \frac{dN_i(u)}{S_{Pi}^2(u)} = \int_0^t \frac{J(u)\sum_{i=1}^n dN_i(u)/S_{Pi}^2(u)}{\{\sum_{i=1}^n Y_i(u)/S_{Pi}(u)\}^2}. \tag{15}$$

The value of (15) is usually greater than (14), which is intuitively reasonable, since the newly proposed estimator (13) also accounts for the unobserved information due to the population censoring.

6. Numerical illustration

A real data example. We compare the estimators described in the previous sections using French data on cancer Waldenstroms macroglobulinemia. The sample consists of 380 patients (226 males, 154 females), with ages ranging from 31 to 97 (mean 71), diagnosed between 1989 and 1997 and followed until 2002. Figure 1a shows the Kaplan-Meier estimate of the observed survival (solid black curve), the expected population survival estimated by the Ederer I method (dotted grey curve), the three traditional relative survival estimators (Ederer I: solid grey; Hakulinen: dashed black; Ederer II: dashed/dotted black) and the newly proposed estimator (dashed black curve). As we can see, the 10 year observed survival is very low and despite the rather old mean age of the patients, the relative survival estimates imply that this is a very severe type of cancer.

[Figure 1 about here.]

Figure 1b compares the bootstrap variance (500 repetitions) to the variance estimated by formula (15). As we can see, both estimates match very well.

Table 1 presents the results of a multivariate additive relative survival model, that assumes Cox-type proportional excess hazard with a piecewise constant baseline excess hazard with breaks at 1, 2, 3, 4, 5 and 10 years. The model was fitted with the maximum likelihood method (Estève et al., 1990). We observe a strong effect of age, but no significant effect of sex and diagnosis year (Table 1). Coefficients for the baseline are not shown.

[Table 1 about here.]

Since the effect of age is in the same direction as in the population tables, individuals with high hazard λ_E get a smaller weight in (6) than in (4) and the observable net survival is above the true net survival curve in Figure 1. Similarly, the weights for those with high population hazard in the second term on the right of (6) are smaller than in (9) and so the Ederer I method overestimates net survival even more than the Ederer II curve. The Ederer I and Hakulinen estimates practically coincide implying there is no important informative censoring due to heterogeneity of potential follow up time.

Additionally, the multivariate model was used to predict the survival curve for each individual, the average over these curves is plotted with a dashed grey curve in Figure 1a. Apart from the fact that it is calculated only at a few time-points (a piece-wise constant model), it coincides with the curve given by our proposed method. This is in accordance with checking the model fit ($p = 0.747$ using the Stare et al. (2005) method).

A simulation study. We continue our illustration by a simulation, with covariate values and simulation parameters defined to mimic the above described real data ($n = 380$). In terms of excess hazard, we consider two situations: with and without dependence on demographic covariates (coefficient for age 0 or 0.05 per year). Times T_{Ei} and T_{Pi} are generated separately, so that both the net survival and the observable net survival can be estimated directly (see Figures 2a and 2c for the average of these curves). These estimates are then compared to the results of the relative survival estimators in Figures 2b and 2d. As we can see, all estimators match between Figures 2a and 2b, since the excess hazard does not depend on demographic covariates. On the other hand, the curves in 2c and 2d differ, with the Ederer II estimate matching the observable net survival, and the newly proposed estimate matching the net survival.

[Figure 2 about here.]

All the relative survival estimators described in this paper are part of the `relsurv` package (Pohar and Stare, 2006) in R (R Development Core Team, 2009) and can be accessed by function `rs.surv` (Ederer I, Ederer II, Hakulinen, new proposal) or `rsadd` (maximum likelihood estimate, `rs.surv.rsadd` for averaging the individual predictions).

7. Discussion and conclusions

It is widely believed that relative survival ratio and net survival represent the same quantity. Whereas this holds when the excess rate does not depend on the demographic variables, it is far from being true in the most usual situation, when this dependence exists. The gap between the two concepts may be large, especially because the excess hazard is almost always highly associated with age at diagnosis. While net survival is by definition the average of individual net survival curves, and as such a survival function, relative survival ratio is a weighted average of individual survival with weights changing with time since diagnosis. Since weights are proportional to the population survival of each individual, this latter measure approaches the net survival of those who have the best population survival. This has been known for many years (Hakulinen, 1977; Buckley, 1984; Estève et al., 1990) but surprisingly it did not prevent the widely accepted belief that the two measures were the same. One of the consequences of this property is the inconsistency of the overall relative survival ratio and its age standardized version which is usually much lower, even if the weights were taken as the weights of the study population (Brenner et al., 2004).

A less often quoted difficulty is that the observable net survival does not have a clear interpretation as a survival function. This means that the Kaplan-Meier estimator in the cause specific setting (with the population deaths considered as censoring) cannot be recommended. When regarded as an estimate of the net survival this estimator is biased because the cancer and non cancer mortality share the influence of the same demographic covariates. The other main variables that have a strong influence on cancer deaths, for example stage, do not

influence non-cancer deaths. Given this fact, it follows, as we have shown, that a consistent estimator may be obtained with an adjustment for the demographic covariates only, either in the context of the cause-specific setting or in the context of a multivariate model for net survival. This same result may be obtained in a more straightforward way with our proposed weighted estimators, either in the relative survival setting or the cause specific setting.

Whereas relative survival ratio is a simple concept and an observable quantity, which has a natural estimator with the Ederer I or Hakulinen method, the net survival is not observable and requires the definition of the additive model. In addition, to make the net survival estimable some hypothesis on the value of the death rate from "other causes" is required. Taking this value from the life tables of the underlying population, we implicitly assume that the death rate from all causes in the patient population is larger than the "control" population death rate. If this assumption is not met the above construction fails and suggests that the life tables do not describe adequately the death rate from other causes in the study population. Such a scenario can be easily recognized by an increasing curve estimated by our newly proposed estimator and will result in convergence problems or biased results with most model fitting approaches.

We have shown that if non cancer deaths are correctly evaluated by the life table of the population, The Ederer II method and the cause specific method estimate the same population value and have equal bias. Our proposed estimators obtained in weighting the individual observation with their population survival provide two unbiased estimators of net survival. When a cause of death has been assigned to each patient and is known from the investigator, it is now possible to make a clean comparison between the two estimates. On the contrary, when comparison between relative and cause-specific survival has been made in the past, it has been contaminated by the bias of different magnitudes present in the two measures.

From a practical point of view, in the usual situation where cause of death is not known,

it is advisable to use the newly proposed method to estimate net survival. It is a method to measure cancer survival after elimination of the influence of other causes of death in large scale multinational survival studies as EUROCARE or CONCORD (Sant et al., 2009; Coleman et al., 2008), where population survival differs substantially between countries. Our measure can be standardized in the traditional manner without the inconsistencies shown by the relative survival ratio, since the net survival of the sample is the average of individual net survival probabilities. This does not preclude the use of multivariate modeling in studies with more precise data on prognostic variables available. A further advantage of our proposal is the consistency of the two approaches since it will estimate the same quantity as an adequate multivariate model, i.e. the "net survival".

We have not addressed the common situation of non-homogeneous potential follow-up time in cancer data. Since such a situation implies informative censoring, it will inevitably lead to bias in all the estimators except a well specified multivariate model. This difficulty was observed by Hakulinen (1982), who realized that the numerator of the relative survival ratio was biased and attempted to diminish the bias of the ratio by introducing a similar bias also in the denominator. While this works well in practice, the development of estimators that would truly remove the bias when estimating either the relative survival ratio or the net survival could be carried out along the lines of our weighted estimator and remains a task for the future.

Acknowledgements

The authors thank the co-editor for many important comments, suggestions and remarks which helped to improve the paper considerably.

References

Aalen, O. O., Borgan, O., and Gjessing, H. K. (2008). *Survival and event history analysis*. Springer-Verlag, New York.

Andersen, P. K. and Vaeth, M. (1989). Simple parametric and nonparametric models for excess and relative mortality. *Biometrics* **45,** 523–535.

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

Buckley, J. D. (1984). Additive and multiplicative models for relative survival rates. *Biometrics* **40,** 51–62.

Coleman, M. P., Quaresma, M., Berrino, F., Lutz, J. M., De Angelis, R., Capocaccia, R., Baili, P., Rachet, B., Gatta, G., Hakulinen, T., Micheli, A., Sant, M., Weir, H. K., Elwood, M. J., Tsukuma, H., Koifman, S., Silva, G. A., Francisci, S., Santaquilani, M., Verdecchia, A., Storm, H. H., and Young, J. L. (2008). Cancer survival in five continents: a worldwide population-based study. *Lancet Oncology* **9,** 730–756.

Cortese, G. and Scheike, T. H. (2008). Dynamic regression hazards models for relative survival. *Statistics in Medicine* **27,** 3563–3548.

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

Ederer, F., Axtell, L. M., and Cutler, S. J. (1961). *The relative survival rate: a statistical methodology*, volume 6, pages 101–121. National Cancer Institute Monograph.

Estève, J., Benhamou, E., Croasdale, M., and Raymond, M. (1990). Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* **9,** 529–538.

Hakulinen, T. (1977). On long-term relative survival rates. *Journal of Chronological Disease* **30,** 431–443.

Hakulinen, T. (1982). Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics* **38,** 933–942.

Hakulinen, T. and Tenkanen, L. (1987). Regression analysis of relative survival rates. *Journal of the Royal Statistical Society — Series C* **36,** 309–317.

Marubini, E. and Valsecchi, M. G. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley.

Nelson, C., Lambert, P. C., Squire, I. B., and Jones, D. R. (2007). Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* **26,** 5486–98.

Pohar, M. and Stare, J. (2006). Relative survival analysis in R. *Computer Methods and Programs in Biomedicine* **81,** 272–278.

Pohar Perme, M., Henderson, R., and Stare, J. (2009). An approach to estimation in relative survival regression. *Biostatistics* **10,** 136–146.

Pokhrel, A. and 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,** 642–647.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the American Statistical Association - Biopharmaceutical Section*, pages 24–33.

Sant, M., Allemani, C., Santaquilani, M., Knijn, A., Marchesi, F., and Capocaccia, R. (2009). EUROCARE-4. survival of cancer patients diagnosed in 1995-1999. results and commentary. *European Journal of Cancer* **45,** 931–991.

Sasieni, P. D. (1996). Proportional excess hazards. *Biometrika* **83,** 127–141.

Satten, G. A., Datta, S., and Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters* **54,** 397–403.

Stare, J., Pohar, M., and Henderson, R. (2005). Goodness of fit of relative survival models. *Statistics in Medicine* **24,** 3911–3925.

Tsiatis, A. A. (2005). *Competing Risks*. John Wiley and Sons, Ltd.

Zahl, P. H. and Aalen, O. O. (1998). Adjusting and comparing survival curves by means of an additive risk model. *Lifetime Data Analysis* **4,** 149–168.

Appendix: The derivation of the newly proposed estimate and its variance

We give here the informal derivation of the newly proposed estimator and its variance that follows closely that for the Nelson-Aalen estimator (see e.g. Aalen et al. (2008)). Using the notation described in the paper, we can write

$$dM_i(t) = dN_i(t) - Y_i(t)d\Lambda_{Pi}(t) - Y_i(t)d\Lambda_E(t).$$

Dividing by the individual population survival curves S_{Pi}

$$\frac{dM_i(t)}{S_{Pi}(t)} = \frac{dN_i(t)}{S_{Pi}(t)} - \frac{Y_i(t)}{S_{Pi}(t)}d\Lambda_{Pi}(t) - \frac{Y_i(t)}{S_{Pi}(t)}d\Lambda_E(t)$$

and using w to denote the weighted processes we get

$$dM_i^w(t) = dN_i^w(t) - Y_i^w(t)d\Lambda_{Pi}(t) - Y_i^w(t)d\Lambda_E(t) \tag{A.1}$$

Summing over individuals and integrating we have

$$M^w(t) = N^w(t) - \int_0^t \sum_{i=1}^n Y_i^w(u)d\Lambda_{Pi}(u) - \int_0^t Y^w(u)d\Lambda_E(u),$$

Note that since S_{Pi} is a predictable process, the process M^w is a martingale with respect to $\mathcal{F}_t = \sigma\{N_i(u), Y_i(u+), S_{Pi}(u+) : 0 \le u \le t, i = 1, \ldots, n\}$. Introducing the indicator function $J(t) = I\{Y(t) > 0\}$ to avoid division by zero ($J(t)/Y(t)$ is interpreted as 0 whenever $Y(t)$ is zero) and summing over individuals, equation (A.1) gives

$$\frac{J(t)}{Y^w(t)}dN^w(t) - \frac{J(t)\sum_{i=1}^n Y_i^w(t)d\Lambda_{Pi}(t)}{Y^w(t)} = J(t)d\Lambda_E(t) + \frac{J(t)}{Y^w(t)}dM^w(t)$$

Integrating the above formula, we thus derive the proposed estimator

$$\widehat{\Lambda}_E(t) = \int_0^t \frac{J(u)}{Y^w(u)} dN^w(u) - \int_0^t \frac{J(u)\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)}.$$

Denoting $\Lambda_E^\star(t) = \int_0^t J(u) d\Lambda_E(u)$, we can see that

$$\begin{aligned} \widehat{\Lambda}_E(t) - \Lambda_E^\star(t) &= \int_0^t \frac{J(u)}{Y^w(u)} dN^w(u) - \int_0^t \frac{J(u)\sum_{i=1}^n Y_i^w(u) d\Lambda_{Pi}(u)}{Y^w(u)} - \int_0^t J(u)\lambda_E(u) du \\ &= \int_0^t \frac{J(u)}{Y^w(u)} dM^w(u), \end{aligned} \tag{A.2}$$

so that $\widehat{\Lambda}_E(t)$ is an unbiased estimator of $\Lambda_E^\star(t)$.

We now use the rules for stochastic integrals (see e.g. Aalen et al. (2008)) to derive a variance estimator. The optional variation process of the martingale in (A.2) equals

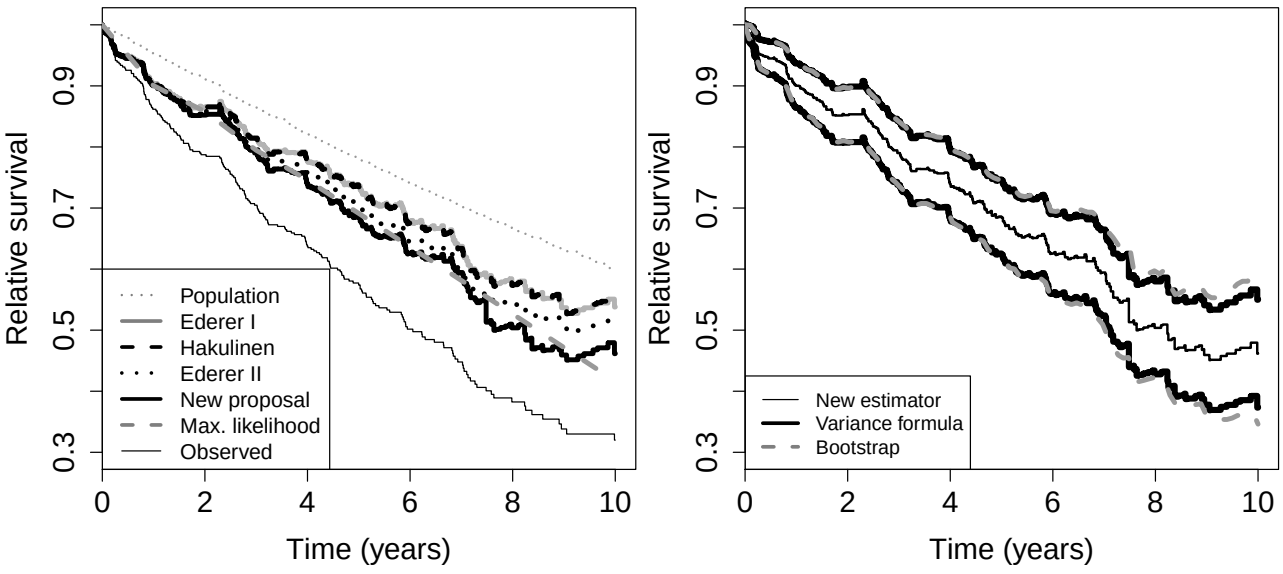
$$[\widehat{\Lambda}_E - \Lambda_E^\star](t) = \int_0^t \frac{J(u)}{Y^w(u)^2} \sum_{i=1}^n \frac{dN_i(u)}{S_{Pi}^2(u)}$$

and since

$$\mathrm{Var}\{\widehat{\Lambda}_\mathrm{E}(\mathrm{t}) - \Lambda_\mathrm{E}^\star(\mathrm{t})\} = \mathrm{E}[\widehat{\Lambda}_\mathrm{E} - \Lambda_\mathrm{E}^\star](\mathrm{t})$$

the variance can be estimated by

$$\widehat{\sigma}^2(t) = \int_0^t \frac{J(u)}{Y^w(u)^2} \sum_{i=1}^n \frac{dN_i(u)}{S_{Pi}^2(u)} = \int_0^t \frac{J(u)\sum_{i=1}^n dN_i(u)/S_{Pi}^2(u)}{\{\sum_{i=1}^n Y_i(u)/S_{Pi}(u)\}^2}.$$



(a) (b)

Figure 1. *(a) Comparison of the estimators in the relative survival setting; (b) comparison of the variance formula and the bootstrap variance for the newly proposed estimator.*

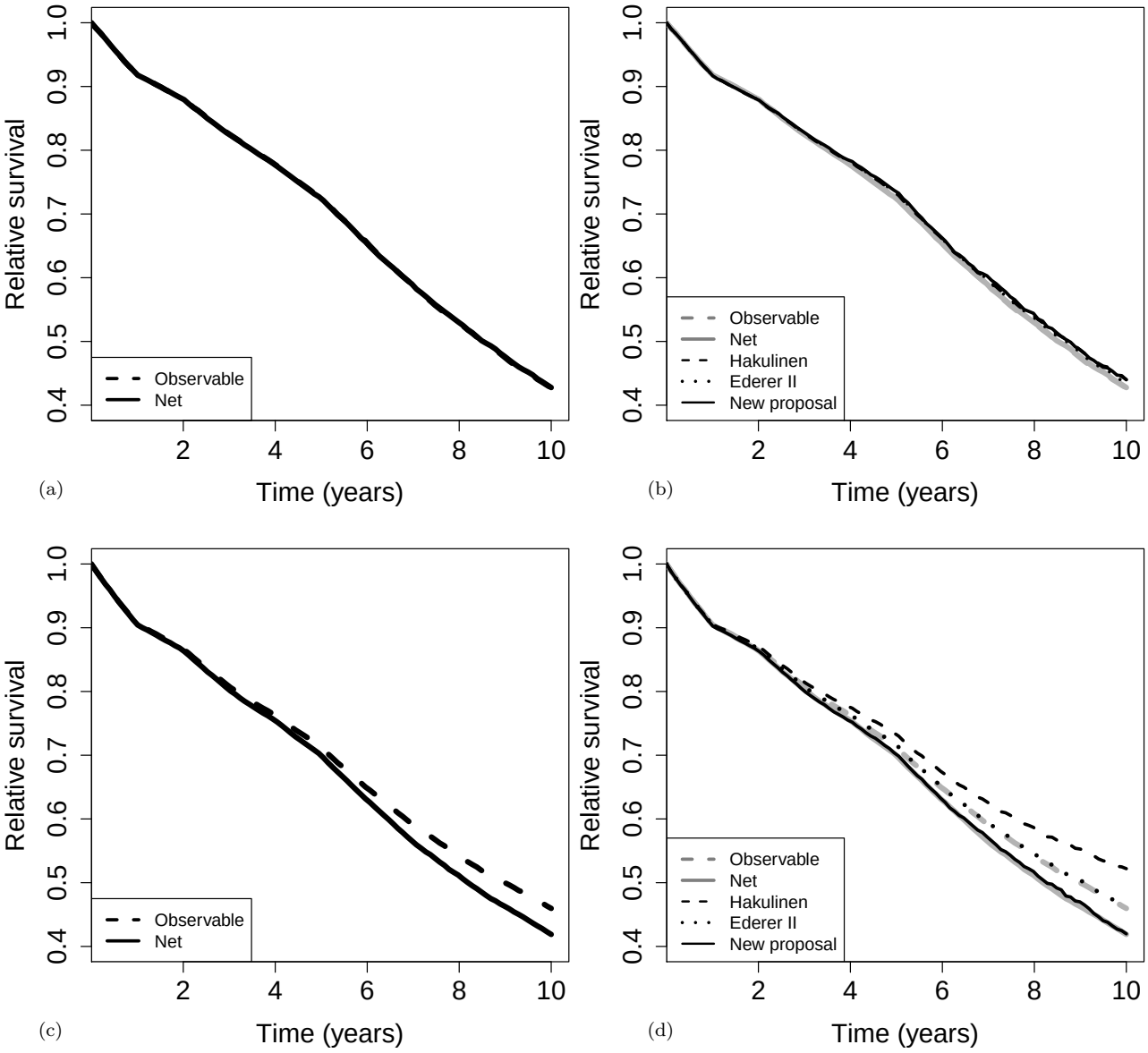


Figure 2. *Comparison of the estimated values: (a) and (b) no effect of covariates on λ_E; (c) and (d) effect of age (coefficient for age* 0.05 *per year). See text for further detail.*

Table 1
The covariate effects.

	Estimate	Std. Error	z value	p
age (per year)	0.054	0.012	4.653	< 0.001
sex (female)	-0.323	0.226	-1.428	0.153
diag. year (per year)	0.077	0.049	1.572	0.116