Direct modelling of the crude probability of cancer death and the number of life-years lost due to cancer without needing the cause of death: a pseudo-observation approach in the relative survival setting

Journal:	*Biostatistics*
Manuscript ID	BIOSTS-18312.R2
Manuscript Type:	Article
Date Submitted by the Author:	19-Mar-2020
Complete List of Authors:	Kipourou, Dimitra-Kleio; London School of Hygiene and Tropical Medicine, pohar perme, maja; University of Ljubljana, Department of Biostatistics and Medical Informatics Rachet, Bernard; London School of Hygiene and Tropical Medicine BELOT, AURELIEN; London School of Hygiene and Tropical Medicine
Keywords:	competing risks, relative survival, pseudo-observations, crude probability of death, number of life years lost

Biostatistics (2020), **0**, 0, *pp.* 1–33
doi:10.1093/biostatistics/main

Direct modelling of the crude probability of cancer death and the number of life-years lost due to cancer without needing the cause of death: a pseudo-observation approach in the relative survival setting

DIMITRA-KLEIO KIPOUROU[a*], MAJA POHAR PERME[b], BERNARD RACHET[a],

AURELIEN BELOT[a]

[a] *Cancer Survival Group, Faculty of Epidemiology and Population Health,*
Department of Non-Communicable Disease Epidemiology,
London School of Hygiene & Tropical Medicine,
London, WC1E 7HT, UK
[b] *Institute for Biostatistics and Medical Informatics,*

Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

dimitra-kleio.kipourou@lshtm.ac.uk

Summary

In population-based cancer studies, net survival is a crucial measure for population comparison purposes. However, alternative measures, namely the crude probability of death (CPr) and the number of life years lost (LYL) due to death according to different causes, are useful as complementary measures for reflecting different dimensions, in terms of prognosis, treatment choice, or development of a control strategy. When the cause of death (COD) information is available, both measures can be estimated in competing risks setting using either cause-specific or subdistribution hazard regression models or with the pseudo-observation approach through direct

[*] To whom correspondence should be addressed.

modelling. We extended the pseudo-observation approach in order to model the CPr and the LYL due to different causes when information on COD is unavailable or unreliable (i.e., in relative survival setting). In a simulation study, we assessed the performance of the proposed approach in estimating regression parameters and examined models with different link functions that can provide an easier interpretation of the parameters. We showed that the pseudo-observation approach performs well for both measures, and we illustrated their use on cervical cancer data from the England population-based cancer registry. The implementation of the method in R software is also provided.

Key words: competing risks, relative survival, pseudo-observations, crude probability of death, number of life years lost

1. Introduction

When aiming to describe the survival experience of a group of individuals, estimating the overall survival is usually of primary interest. However, when the goal is to describe the probabilities of dying from different causes, a further step is required in order to account for competing risks. Competing risks methods aim to identify covariates which not only affect the rate at which specific events occur, but also the probability of occurrence of a specific event over time (Austin and Fine, 2017).

To perform a competing risks analysis with two events, say cancer death and death from other causes, we often rely on the cause of death (COD) information attributed to each individual, assuming that this is available and reliable. Two types of hazards, namely cause-specific and subdistribution hazard, may be used. Unlike cause-specific hazard, subdistribution hazard is useful for estimating covariate effects on the event-specific probability since it "measures the effect of the covariate that can be explained either because there is a direct effect of making the event more or less likely to occur, or due to the indirect effect of influencing the other events to occur" (Dignam and Zhang, 2012). However, this leads to a complicated interpretation thus, working with cause-specific hazard might be preferred even though it does not directly describe the covariate effect on the probabilities (Andersen and Keiding, 2012).

Nevertheless, the use of routinely collected population-based registry data involves additional methodological challenges due to the absence of reliable information on individual COD, calling for methods defined within the *relative survival* setting (Pohar Perme *and others*, 2016). In this setting, the observed mortality hazard is split into two mortality hazards: the expected or population mortality hazard (assumed known and provided by population life tables) and the excess mortality hazard, which is the main quantity of interest. The excess mortality hazard in the relative survival setting is the equivalent of cause-specific (here cancer-specific) hazard in classical competing risks setting. The most frequently used indicator derived from the excess mortality hazard is *net survival* (Pohar Perme *and others*, 2012), which is the probability of surviving when assuming that the cancer under study is the only possible COD. Net survival is of interest when making comparisons between populations since it is independent of the competing risks of death, which may differ between these populations (Allemani *and others*, 2018; De Angelis *and others*, 2014).

Despite the usefulness of net survival, communicating survival statistics is complicated and must involve various indicators, as to reflect different dimensions in terms of prognosis, treatment choice, or development of a control strategy. Towards this direction alternative indicators like (i) the *Crude Probability of Death* (CPr) from a given cause (Cronin and Feuer, 2000; Mariotto *and others*, 2014; Pfeiffer and Gail, 2017), also called cause-specific cumulative incidence function, and (ii) the number of *Life Years Lost* (LYL) due to a given cause (Andersen, 2013), can be used as complementary tools in order to provide a multi-perspective approach (Belot *and others*, 2019).

Crude probabilities can be estimated nonparametrically, using Aalen-Johansen estimator (Satagopan *and others*, 2004; Geskus, 2015), or modelled in cause-specific setting with regression models on the cause-specific hazards (Pfeiffer and Gail, 2017; Kipourou *and others*, 2019) or on the subdistribution hazards (Fine and Gray, 1999; Geskus, 2015; Mozumder *and others*, 2018) or modelled in relative survival setting using regression models on the excess hazard (Lambert *and others*, 2010; Eloranta *and others*, 2013; Charvat *and others*, 2013, 2016). In cause-specific setting, the pseudo-observation approach is another option (Andersen *and others*, 2003; Klein and Andersen, 2005; Andersen and Pohar Perme, 2010) allowing for the direct modelling of probabilities. In relative survival setting CPr can be estimated nonparametrically or indirectly via regression modelling of the excess hazard but not through direct modelling. Similarly, although estimation and modelling of LYL can be implemented in cause-specific setting (Andersen, 2013), modelling them in relative survival setting has yet to be implemented.

The scope of this paper is to present a way of modelling directly the CPr and LYL due to the disease of interest and other causes in relative survival setting (i.e., when COD is not available) according to some covariates of interest. We chose to extend the most general method, i.e., the pseudo-observation method (Andersen *and others*, 2003; Klein and Andersen, 2005; Andersen and Pohar Perme, 2010; Andersen, 2013), which can be applied to both measures. The main idea is based on the fact that when there is censoring we do not always observe the random variable (e.g. time to event). By generating pseudo-observations at specific time points, we replace the whole set of incompletely observed random variables with a complete set of their pseudo-observations. These are later modelled with standard methods like generalised linear models (GLM) or generalised estimating equations (GEE) in order to quantify covariate effects directly on the indicators of interest.

The remainder of the paper is organised as follows: Section 2 provides a general description of the pseudo-observation approach and details how this can be adapted in the relative survival setting in order to model directly the CPr and LYL. In Section 3, we assessed the performance of the method in its ability to estimate the regression parameters of interest and examined models with different link functions using simulations. In Section 4, we applied the new method on population-based cancer registry data of women diagnosed with cervical cancer in England between 2008 and 2010, and discussed the useful interpretation that can be gained from these models. Lastly, Section 5 summarises the results and presents ideas for further research.

2. Methods

2.1 *Pseudo-observations*

The method based on pseudo-observations provides a general framework that enables the direct modelling of a given statistical measure (e.g. survival probability) as a function of some covariates of interest. Pseudo-observations (also called pseudo-values) were first described for multistate models (Andersen *and others*, 2003), and since then many extensions were proposed (e.g. for cause-specific cumulative probabilities within the classical competing risks setting, (Klein and Andersen, 2005; Moreno-Betancur and Latouche, 2013) or for the restricted mean survival time (Andersen *and others*, 2004)). This approach requires the existence of an (approximately) unbiased estimator of the measure of interest (Andersen and Pohar Perme, 2010). While its usefulness goes beyond modelling (as it can be extended to providing goodness-of-fit methods (Andersen and Pohar Perme, 2010; Pavlič *and others*, 2018)), we focus on the modelling part here, and summarize the main steps for their use when analysing time to event data.

For an individual $i = 1, \cdots, n$, let Y_i be independent and identically distributed random variables (e.g. time since diagnosis up to death), and $\boldsymbol{X}_i$ a p-dimensional vector of (time-fixed) covariates. As it is often the case with time to event data analysis, Y_i is not always observed due to censoring. Pseudo-observations are useful when information on Y_i is not available, and our interest lies on modelling the $E\left[f(Y_i)|\boldsymbol{X}_i\right]$ for a given function f.

The main idea of pseudo-observations relies on the fact that even with incomplete (censored) data we can still derive the marginal expectation $E\left[f(Y)\right]$. Assuming that a consistent and (approximately) unbiased estimator $\hat{\theta}$ exists for $\theta = E\left[f(Y)\right]$ (e.g. the Kaplan-Meier estimator for the survival probability, or the Aalen-Johansen estimator for the cause-specific cumulative incidence function (Geskus, 2015)), then the possibly unknown $f(Y_i)$ could be replaced by its pseudo-observation (Andersen and Pohar Perme, 2010).

Pseudo-observations are computed for every individual regardless of the availability of the $f(Y_i)$ at specific times. Thus, the pseudo-observation for $f(Y_i)$ is defined for individual $i = 1, \cdots, n$ at a given time t as

$$\tilde{\theta}_i = n \cdot \hat{\theta} - (n-1) \cdot \hat{\theta}^{-i} \tag{2.1}$$

where $\hat{\theta}$ is the estimator at time t based on the whole sample and $\hat{\theta}^{-i}$ is the estimator at time t based on the sample of size $(n-1)$, obtained by eliminating individual i from the whole sample. Intuitively, the pseudo-observation $\tilde{\theta}_i$ can be seen as the "contribution" of the individual i to the $E\left[f(Y_i)|\boldsymbol{X}_i\right]$, estimated on the basis of the full sample at time t (Andersen and Pohar Perme, 2010).

Pseudo-observations may be calculated at several time points. In this case, the pseudo-observation $\tilde{\boldsymbol{\theta}}_i$ is m-dimensional (i.e., $(\tilde{\boldsymbol{\theta}}_i)_j = \tilde{\theta}_{ij}, \quad j = 1, \cdots, m$) and represents the vector $f(\boldsymbol{Y}_i)$ $(\boldsymbol{Y}_i = (Y_{i1}, \cdots, Y_{im}))$ with entries $f(Y_{ij})$. These pseudo-observations may be used as the outcome variables in a generalised linear regression model in order to derive the covariate effects on the outcome of interest as

$$g\{E[f(Y_{ij})|\boldsymbol{X}_i]\} = \boldsymbol{\alpha}_j + \boldsymbol{\gamma}^\top \boldsymbol{X}_i = \boldsymbol{\beta}^\top \boldsymbol{X}^*_{ij}, \qquad i = 1, \cdots, n \quad j = 1, \cdots, m \tag{2.2}$$

where g is a monotone differentiable link function and $\boldsymbol{X}^*_{ij}$ is a $(m+p$ dimensional) vector including the indicators of the time points and the covariates $\boldsymbol{X}_i$, $\boldsymbol{X}^*_{ij} = (e_j^\top, \boldsymbol{X}_i^\top)^\top$ where e_j is the m-dimensional vector with 1 on the jth entry and 0 otherwise (Andersen and Pohar Perme, 2010). Adding interaction terms (between covariates and time terms) would make $\boldsymbol{X}^*_{ij}$ higher dimensional.

Because the pseudo-observations for a given subject could not be considered as independent random variables, estimating the $(m+p)$ regression parameters $\boldsymbol{\beta}$ is based on the Generalised Estimating Equations (GEE) method (Liang and Zeger, 1986). The estimating equations to be solved are

$$\sum_{i=1}^{n} \left(\frac{\partial}{\partial \boldsymbol{\beta}} g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}^*_i) \right)^\top V_i^{-1} \left\{ \tilde{\boldsymbol{\theta}}_i - g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}^*_i) \right\} = 0 \qquad (2.3)$$

where $g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}^*_i)$ is an m-dimensional vector $(g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}^*_{i1}), \cdots, g^{-1}(\boldsymbol{\beta}^\top \boldsymbol{X}^*_{im}))$ and V_i is a working covariance matrix with a pre-defined structure.

In order for the pseudo-observation approach to work, it has been shown that the censoring should not depend on covariates (Graw *and others*, 2009), alternatively modified pseudo-observations should be applied (Binder *and others*, 2014). For the variance of the estimated regression parameters $\hat{\boldsymbol{\beta}}$, a sandwich estimator could be used (Andersen *and others*, 2003). Even if it has been shown that this might lead to inconsistent and upward biased results (especially in the case of large samples) this has an insignificant impact in practical applications (Jacobsen and Martinussen, 2016; Overgaard *and others*, 2017, 2018).

The user has various choices with respect to the link function g and the structure of the working covariance matrix V. A clever choice of the latter may increase efficiency (Andersen, 2013), but we do not explore this further in this study.

2.2 *The relative survival setting and the excess mortality hazard approach*

The relative survival setting is a specific competing risks setting where, although the COD information is either missing or is unreliable, inference about the event/disease of interest can still be drawn under specific assumptions and conditions (detailed below). In this paper, the disease of interest is a specific cancer and the time scale used for measuring the time to event is the time since cancer diagnosis.

In relative survival setting, we consider two sets of data: i) data on time to death (but without COD information) from a cohort of patients with the specific cancer of interest and, ii) life tables of the general population in which all-cause hazard functions (stratified by some sociodemographic variables z) are available (Pohar Perme *and others*, 2012). The main assumption we make here is that for an individual i, the observed hazard $\lambda_O(t;\boldsymbol{X}_i)$ described by the covariates $\boldsymbol{X}_i$ can be decomposed as the sum of the cancer-specific mortality hazard $\lambda_C(t;X_i)$ and the hazard related to other causes $\lambda_P(t;z_i)$ (with $z_i \subset \boldsymbol{X}_i$):

$$\lambda_O(t;\boldsymbol{X}_i) = \lambda_C(t;\boldsymbol{X}_i) + \lambda_P(t;z_i) \tag{2.4}$$

We further assume that λ_P is equal to all-cause hazard of the general population within levels of z. For this assumption to hold, the following conditions must be met:

- the specific cancer of interest is considered a negligible cause of death in the general population (Ederer, 1961). This is especially true when prevalence is low (i.e., rare cancers and younger age groups), but it might be unreasonable when focusing for example on older people with common cancers (e.g. prostate cancer) or when all cancer sites are combined (Hinchliffe *and others*, 2012; Talbäck and Dickman, 2011).

- the other-cause hazard of the general population is equal to the other-cause hazard of the study population within levels of z. Moreover, within levels of z, the other-cause hazard does not further depend on $\boldsymbol{X}$ nor on any (unmeasured) covariates. This latter condition may not be realistic for some cancers, and an adaptation of the method might be needed (Rubio *and others*, 2019).

In most situations, the minimum set of sociodemographic covariates z stratifying life tables (and therefore λ_P) is sex, age (in 1-year age-group), calendar year and geographical level. In some countries, additional stratifying variables may be available, such as deprivation level or ethnicity.

A discussion of the assumptions and related conditions that should be met for the relative survival setting to be valid can be found in (Pavlič and Pohar Perme, 2018).

2.3 *Measures of interest in the relative survival setting*

2.3.1 *Crude probability of death from a specific cause* In the classical competing risks setting where the COD is available, the (cause k)-specific probability of death $F_k(t)$ (also called cumulative incidence function) represents the probability of dying from cause-k before or at time t, and can be expressed as $F_k(t) = \int_0^t S(u-)\mathrm{d}\Lambda_k(u)$, where S is the all-cause survival function and Λ_k is the cumulative (cause k)-specific hazard.

In the relative survival framework, the crude probability of death from cancer ($F_{\mathrm{C}}(t)$) is expressed as $F_{\mathrm{C}}(t) = \int_0^t S(u-)\mathrm{d}\Lambda_C(u)$ (Cronin and Feuer, 2000; Lambert *and others*, 2010; Charvat *and others*, 2013). It may be estimated using the marginal cancer-specific hazard $\lambda_C(t)$, this latter being defined as the combination of the *individual* cancer-specific hazards, $\lambda_C(t, \boldsymbol{X})$ (see equations 5 and 6 in (Pohar Perme *and others*, 2012), while more details can be found in (Belot *and others*, 2019; Pohar Perme and Pavlic, 2018)). Thus, it holds that

$$\hat{F}_{\mathrm{C}}(t) = \int_0^t \hat{S}(u-)\mathrm{d}\hat{\Lambda}_C(u) \tag{2.5}$$

where $\hat{S}(u-)$ is the Kaplan-Meier estimator of the overall survival and the estimator of the cancer-specific cumulative hazard is calculated as

$$\mathrm{d}\hat{\Lambda}_C(t) = \frac{\mathrm{d}N(t) - \sum_{i=1}^{n} Y_i(t)\mathrm{d}\Lambda_P(t, \mathbf{z}_i)}{Y(t)}$$

Similarly, the crude probability of death from other causes can be estimated as

$$\hat{F}_{\mathrm{P}}(t) = \int_0^t \hat{S}(u-)\mathrm{d}\hat{\Lambda}_P(u) \tag{2.6}$$

where

$$\mathrm{d}\hat{\Lambda}_P(t) = \frac{\sum_{i=1}^{n} Y_i(t)\mathrm{d}\Lambda_P(t, \mathbf{z}_i)}{Y(t)}$$

In both formulae, $\mathrm{d}\Lambda_P$ is obtained through λ_P, which is the population mortality hazard that an individual i with covariates $\mathbf{z}_i$, $i = 1, \cdots, n$, is exposed to at time t. $N(t)$ and $Y(t)$ are counting processes, where $N(t)$ is the number of individuals who have experienced an event of any type in $[0, t]$, and $Y(t)$ is the number of individuals who are still at risk at time t, obtained as the sum of indicators whether a person is still at risk, $Y(t) = \sum Y_i(t)$ (Klein and Andersen, 2005; Andersen and Pohar Perme, 2010; Pohar Perme and Pavlic, 2018).

This method of estimation is already provided in the R-package `relsurv` (Pohar Perme, 2018).

2.3.2 *Number of life years lost due to a specific cause* The expected LYL due to a specific cause (for a given time window) is a useful complementary indicator (Andersen, 2013), allowing for an easier interpretation of the results, which is expressed with time units. In clinical settings, this indicator provides an interesting insight on prognosis, treatment choice, or the development of a control strategy.

Without distinguishing death from different causes, the LYL before time τ (compared to an immortal cohort (Andersen, 2013), *ie* where nobody dies before time τ), may be expressed as

$$L(0, \tau) = \tau - \int_0^{\tau} S(u)\mathrm{d}u$$

The total LYL can be further decomposed according to COD in the classical competing risks setting as $L_k(0, \tau) = \int_0^{\tau} F_k(u)\mathrm{d}u$ where $F_k(t)$ is the cause k-specific cumulative probability of death (Andersen, 2013). Therefore, following the same analogy as before, this decomposition can be extended to the relative survival setting for the LYL due to cancer L_C and due to other-cause L_P (Belot *and others*, 2019):

$$L_C(0, \tau) = \int_0^{\tau} F_C(u)\mathrm{d}u, \quad L_P(0, \tau) = \int_0^{\tau} F_P(u)\mathrm{d}u \tag{2.7}$$

Finally, by plugging in the estimators (2.5) and (2.6) in the equation (2.7) we can estimate the $\hat{L}_C(0,\tau)$ and $\hat{L}_P(0,\tau)$, respectively.

2.4 *Pseudo-observations in the relative survival setting for estimating covariates effects on the CPr and the LYL due to different causes*

The pseudo-observation for the CPr due to cancer for an individual i at time t, $\tilde{F}_{C,it}$ is calculated (based on the equations (2.1) and (2.5)) as

$$\tilde{F}_{C,it} = n \cdot \hat{F}_C(t) - (n-1) \cdot \hat{F}_C^{-i}(t) \tag{2.8}$$

This pseudo-observation is defined at a particular timepoint, and for regression modelling it was advised to calculate the pseudo-observations at m between 5 and 10 different timepoints, which can be either equally spread or chosen based on quantiles of the overall survival time distribution (Klein and Andersen, 2005). The pseudo-observations for the CPr of death due to other causes are defined in the same way.

For the LYL due to cancer $L_{C,i}(0,\tau)$ (resp. other cause, $L_{P,i}(0,\tau)$), we compute only m=1 pseudo-observation at time τ for each individual (based on the equations (2.1), (2.7)) as

$$\tilde{L}_{C,i}(0,\tau) = n \cdot \hat{L}_C(0,\tau) - (n-1) \cdot \hat{L}_C(0,\tau)^{-i} \tag{2.9}$$

For both indicators, after calculating these pseudo-observations we generate a new dataset in which every individual is assigned with m pseudo-observations (corresponding to the m timepoints), which later will be used as the main outcome in a regression model (Andersen *and others*, 2003). A GEE model of the form $g(E[Y|\boldsymbol{X}_i]) = \boldsymbol{\beta}^\top \boldsymbol{X}^*_{ij}$ is typically used, where g is a link function, $\boldsymbol{\beta}$ is the corresponding vector of $m+p$ regression parameters, and $\boldsymbol{X}^*_{ij}$ is a vector including the covariates for the individual i ($\boldsymbol{X}_i$) as well as the intercept and the indicator functions of the $(m-1)$ remaining timepoints.

2.4.1 *User choices: link function and working covariance matrix*

Interpretation of regression coefficients varies according to the link function used. For the CPr, most common g link functions are the *cloglog*, *log*, and *identity*.

A *cloglog* link function on $F_C(t)$, defined as $\log(-\log(1 - F_C))$ leads to similar regression coefficient estimates to those obtained with Fine & Gray model (Fine and Gray, 1999). In this case, the $\exp(\boldsymbol{\beta})$ is a hazard ratio which is related to the subdistribution hazard, i.e., the instantaneous rate of failure per time unit from cause j among those who are either alive or have had a competing event at time t. Due to the complicated nature of this type of hazard ratios, the regression coefficients are interpreted in a qualitative (higher or lower than 1) rather than quantitative way (Andersen *and others*, 2012). Nonetheless, a test of statistical significance of a subdistribution hazard ratio provides a test of the covariate effect on the CPr (Austin and Fine, 2017).

A *log* link function gives regression coefficients with less complicated interpretation. The $\exp(\boldsymbol{\beta})$ obtained from a model with log link function gives an estimate of the relative risks (Overgaard *and others*, 2015) allowing for quantitative interpretations. However, constraining probabilities between [0,1] might be problematic in situations with high absolute risks or when extrapolating outside the data range (Lambert *and others*, 2017).

Additionally, an *identity* link function can be applied to CPr leading to regression coefficients that are interpreted as risk differences (Klein, 2006; Hansen *and others*, 2014). The identity link function is usually the link function of choice for models on LYL as well. In this case, the interpretation shows the additional life years that are lost due to a given cause. In both cases though, results might go beyond the admissible range which is set for each indicator and thus, one must be careful of predictions outside the observed limits.

The *logit* link function (not explored here) would be another option giving also convenient interpretations i.e., odds ratios. This choice also suffers from certain drawbacks like for example numerical instabilities for small values of time t (Gerds *and others*, 2012).

We account for the correlation in the pseudo-observation data through the use of a specific structure of the working covariance matrix (Pekár and Brabec, 2018). The choice of a covariance matrix structure might vary between independence, exchangeable, autoregressive and unstructured, although it is suggested that even the independence working covariance matrix is adequate (Klein and Andersen, 2005).

3. Simulation study

In this study, we conjecture that the pseudo-observation approach for the relative survival setting will work in a similar way as in the classical competing risks setting and GEE would be a reasonable approach to yield both regression parameter and variance estimates. With a simulation study we examine the validity of the method in practice. Simulations were performed in order to evaluate the frequentist properties of the proposed method based on pseudo-observations, in its ability to estimate regression parameters of covariates associated to CPr and LYL due to death from cancer and from other causes.

3.1 *Data generation and simulation design*

We simulated $n_{\text{sim}} = 500$ datasets with sample size of $N = \{300, 1000\}$. Each individual was assigned with a vector of three covariates which includes information about sex, year of diagnosis, and age at diagnosis. Sex was simulated as binary drawn from a Bernoulli distribution with probability 0.5. Year of diagnosis was simulated as a continuous variable and sampled from a uniform distribution, ranging from 2000 to 2003. Age at diagnosis was simulated as a continuous variable by first selecting an age class according to predefined probabilities (0.25 for age class [30,65), 0.35 for age class [65,75) and 0.4 for age class [75,80)) and then sampling from a class-specific uniform distribution (Belot *and others*, 2010).

This scenario tried to mimic what could be observed in real situations for colon cancer patients. We used a Generalised Weibull distribution with parameters (κ, ρ, α) to model the subdistribution hazard (SDH). For individual i, the SDH related to cancer γ_C was defined as

$$\gamma_C(t, Age_i, Sex_i, Year_i) = \gamma_0(t) \exp\{\beta_{Age} Age_i + \beta_{Sex} Sex_i + \beta_{Year} Year_i\}$$

where

$$\gamma_0(t) = \frac{\kappa \rho^\kappa t^{\kappa-1}}{1 + \frac{(\rho t)^\kappa}{\alpha}}$$

The parameters used here, namely $\{\kappa, \rho, \alpha\}$, for the baseline hazard were set to $\{2, 1.6, 0.05\}$. The values used for the covariate regression parameters were $\beta_{Age} = 0.2$ (for 1 year increase), $\beta_{Sex} = 0.3$, and $\beta_{Year} = 0$, accounting for different strength in effect sizes; a very strong effect (age), a weak effect (sex), and a null effect (year). In this way, simulations include the most common covariates used in relative survival analyses.

We obtained the expected mortality λ_P from UK life tables based on some demographic characteristics, namely year, age and sex (Danieli *and others*, 2012). The λ_P changed annually for a given age and sex and remains constant during a year hence, following a piecewise exponential distribution.

Using γ_C and λ_P, we obtained the cancer specific hazard λ_C by adapting the approach described in (Haller and Ulm, 2014). The individual survival time (from any cause) was obtained using the inverse probability transform method (Bender *and others*, 2005; Belot *and others*, 2010). More information on the simulation algorithm are provided in Section 1 of Supplementary Material A.

We set the administrative censoring time (C) at 10 years and allowed for a separate distribution to account for drop-outs, which followed an exponential distribution $(\lambda_d = 0.035)$. This results in approximately 8% loss to follow-up, while the total amount of censoring in each dataset was on average around 42%. A vital status indicator δ was created, $\delta = 0$ for individual censored at T and $\delta = 1$ for those being dead at time T (irrespective of COD).

3.2 *Analysis of simulated data*

For CPr from cancer and other causes, we tested three GEE models for the pseudo-observations: (a) a model with *log* link function, (b) a model with *cloglog* link function, and (c) a model with *identity* link function. All models were assuming independence working correlation and included the explanatory variables age at diagnosis, sex, and year of diagnosis.

To model LYL within 10 years from death caused by cancer or other causes, we fitted a GEE model with *identity* link function, explanatory variables: age at diagnosis, sex, and year of diagnosis, and independence covariance structure.

We calculated the following performance measures:

1. bias, defined as the difference between the average of the $n_{sim} = 500$ estimated values and the true value β_0^*: $\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\beta}_i - \beta_0^*$,
2. empirical standard error $\sqrt{\frac{1}{n_{sim}-1} \sum_{i=1}^{n_{sim}} (\hat{\beta}_i - \bar{\beta})^2}$ where $\bar{\beta} = \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \hat{\beta}_i$,
3. model standard error $\sqrt{\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} [\hat{Var}(\hat{\beta}_i)]}$,
4. root mean squared error $\sqrt{\frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} (\hat{\beta}_i - \beta_0^*)^2}$, and
5. the coverage which is the proportion of samples in which the 95% confidence interval included β_0^*.

Having simulated using the subdistribution hazard, our generated data followed the specified model with the cloglog link function. For the other two link functions and the other causes with cloglog (where real population hazards were taken), the performance was assessed indirectly with the least false parameters (LFP) (Hjort, 1992; Beyersmann *and others*, 2009). The LFP were obtained after applying the same models described previously to a dataset of 100,000 individuals, which was generated using the same simulation algorithm but without considering any drop outs (for more details please see Section 2 of Supplementary Material A). Both true and LFP were available for the cloglog, so this case allowed us to evaluate the sensibility of the LFP as proxies of the true values. The LFP for model (b) for the cancer case were $(0.199, 0.299, 0.005)$ whereas the true (simulated) were $(0.2, 0.3, 0)$, validating this way of comparison.

Our computations were performed in R 3.2.0. We used the nonparametric method for the CPr provided by the R-package `relsurv` (version 2.1.1, function `cmp.rel` (Pohar Perme, 2018)), while GEE models were fitted with the R-package `geepack` (version 3.2.5, function `geese`).

3.3 *Simulation results*

3.3.1 *CPr of death from colon cancer and other causes* Results shown in Table 1 suggested that regardless of the link function used, the regression parameter estimates of the covariate effects were almost unbiased with most of the coverage probabilities lying within the acceptable coverage range ($[0.931, 0.969]$) for all parameter estimates and for any cause (cancer or other causes). Results were similar for both sample sizes although, for model (c) results seem to be slightly better when $N = 1000$ due to a smaller bias in the larger sample size. In general, standard error was found to be adequately estimated with the models based on how close the empirical standard errors compared to model standard errors are. RMSEs were also reasonably low proving also good model performance.

The only exception to that is the regression parameter estimates in model (c) in the case of age (for cancer) and year (for both causes) when $N = 300$. In all cases, standard error seemed to be well estimated thus, indicating that the bias in the estimator should be probably the reason for the problematic coverage probability. A different choice of working correlation structure would change both the regression parameter estimate and its variance, leading to a possibly better coverage probability, while model misspecification might be an additional issue which may be considered.

3.3.2 *Life years lost* The regression parameters were well estimated when modelling the number of LYL due to each cause, with a very small bias and a good coverage (see Table 2). Only exception to that was the estimated regression parameter for the effect of sex and year in the case of other causes when $N = 1000$. An overestimation of the standard error by the model might have inflated the coverage probability in case of sex, while bias seems to be the source of problem in the case of year. Another specification of the model including a change of the working covariance matrix would be additional things to consider.

4. Illustrative example

We illustrated our approach using a dataset of 7351 women diagnosed in England with cervical cancer between 2008 and 2010, obtained from the national population-based cancer registry. We limited the sample to those aged between 15 and 89 years, the end of follow-up was set at the 31st of December of 2015 and all individuals had a minimum potential follow-up of 5 years. In this dataset, 2255 (30.7%) deaths were observed (whatever the causes, as the exact COD was not available) while 186 (2.5%) were lost to follow-up. We applied the nonparametric method and the pseudo-observations approach defined in the relative survival setting, and we used the UK life tables, stratified by sex, age, calendar year, government office region and deprivation quintiles.

The covariates of interest for studying their association with the crude probabilities of death or with the number of LYL due to each cause were: age at diagnosis defined as a continuous variable, and the deprivation quintiles. For the latter, patients were categorized in 5 socioeconomic groups (from the least deprived group, level 1, to the most deprived group, level 5) using national categories of the income domain of the Index of Multiple Deprivation score (IMD 2004) which is a score defined at the lower super output area level in England.

In this illustration, our aim was to obtain and interpret the regression parameter estimates which quantify the effect of covariates on CPr and LYL due to cancer and other causes. We show

Table 1. Simulation results: performance measures of regression parameter estimated using pseudo-observation and 3 models for the crude probabilities of death from cancer and from other causes; model (a) assumed a *log* link function, model (b) assumed a *cloglog* link function, and model (c) assumed an *identity* link function. In all models, the independence working covariance structure was used. The explanatory variables in all models were age at diagnosis, sex, and year of diagnosis. Results based on 500 simulated datasets with different sample sizes ($N = 300, 1000$).

Model	**Cause**	**Covariate**	**Least**	$\hat{\beta}$		**Bias** (×1000)		**empSE**†		**ModSE**†		**RMSE**†		**Coverage**‡	
			false	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000
(a)	*Cancer*	Age	0.163	0.174	0.165	10.649	1.869	0.08	0.038	0.075	0.04	0.081	0.039	0.932	0.956
		Sex	0.24	0.263	0.249	22.386	8.587	0.218	0.118	0.21	0.115	0.219	0.118	0.948	0.944
		Year	0.004	-0.006	0	-10.162	-3.997	0.109	0.056	0.102	0.055	0.109	0.056	0.936	0.946
	Other causes	Age	0.693	0.709	0.694	16.23	1.664	0.105	0.056	0.111	0.058	0.107	0.056	0.952	0.944
		Sex	0.158	0.163	0.148	4.38	-10.76	0.164	0.079	0.158	0.083	0.164	0.08	0.932	0.966
		Year	-0.016	-0.012	-0.016	3.887	0.341	0.093	0.049	0.088	0.047	0.093	0.049	0.936	0.936
(b)	*Cancer*	Age	0.2*	0.212	0.202	12.071	1.623	0.098	0.047	0.09	0.048	0.098	0.047	0.932	0.95
		Sex	0.3*	0.325	0.309	25.02	8.801	0.261	0.143	0.253	0.139	0.262	0.143	0.948	0.948
		Year	0*	-0.007	0	-7.334	-0.037	0.137	0.07	0.13	0.07	0.137	0.07	0.936	0.94
	Other causes	Age	0.793	0.81	0.794	17.059	1.073	0.125	0.066	0.129	0.068	0.126	0.066	0.948	0.942
		Sex	0.194	0.195	0.184	1.128	-10.17	0.181	0.087	0.175	0.092	0.181	0.088	0.936	0.966
		Year	-0.019	-0.015	-0.019	4.625	0.758	0.104	0.055	0.098	0.052	0.104	0.055	0.938	0.938
(c)	*Cancer*	Age	0.04	0.04	0.04	-0.326	-0.259	0.015	0.008	0.015	0.008	0.015	0.008	0.926	0.946
		Sex	0.064	0.066	0.064	2.358	0.682	0.048	0.028	0.049	0.027	0.048	0.028	0.948	0.942
		Year	0.002	0	0	-2.068	-1.637	0.029	0.015	0.028	0.015	0.029	0.015	0.924	0.958
	Other causes	Age	0.037	0.037	0.038	-0.302	0.789	0.003	0.002	0.003	0.002	0.003	0.002	0.97	0.964
		Sex	0.019	0.022	0.019	3.329	0.795	0.009	0.005	0.009	0.005	0.009	0.005	0.932	0.962
		Year	-0.002	0.001	-0.003	2.718	-0.835	0.006	0.003	0.006	0.003	0.006	0.003	0.912	0.928

∗true values

† empSE: empirical standard error; ModSE: model standard error; RMSE: root mean squared error

‡ Acceptable coverage range is calculated as $0.95 \pm z_\alpha \sqrt{\frac{0.95 \cdot 0.05}{500}} = [0.931, 0.969]$

Table 2. Simulation results: performance measures of regression parameter estimated using pseudo-observation and a model with *identity* link function and an independence working covariance matrix for the number of Life Years Lost due to cancer and due to other causes. The explanatory variables were age at diagnosis, sex, and year of diagnosis. Results based on 500 simulated datasets with different sample sizes (N=300,1000).

Cause	**Covariate**	**Least**	$\hat{\beta}$		**Bias** (×1000)		**empSE**[†]		**ModSE**[†]		**RMSE**[†]		**Coverage**[‡]	
		false	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000	N=300	N=1000
Cancer	Age	0.553	0.542	0.55	-10.029	-2.559	0.179	0.097	0.177	0.097	0.18	0.097	0.948	0.948
	Sex	0.749	0.8	0.777	50.949	28.345	0.569	0.318	0.573	0.319	0.572	0.319	0.954	0.95
	Year	-0.002	-0.002	-0.003	-0.133	-1.913	0.322	0.17	0.321	0.172	0.322	0.17	0.954	0.956
Other causes	Age	0.49	0.489	0.502	-0.737	12.319	0.045	0.025	0.049	0.028	0.045	0.027	0.968	0.964
	Sex	0.223	0.273	0.235	49.817	12.618	0.134	0.07	0.139	0.077	0.143	0.071	0.934	0.974
	Year	-0.021	0.011	-0.037	31.802	-15.82	0.084	0.047	0.085	0.047	0.09	0.05	0.932	0.92

† empSE: empirical standard error; ModSE: model standard error; RMSE: root mean squared error

‡ Acceptable coverage range is calculated as $0.95 \pm z_\alpha\sqrt{\frac{0.95 \cdot 0.05}{500}} = [0.931, 0.969]$

how the interpretation changes based on the choice of link function, while we demonstrate in practice the advantages and limitations that were described in Section 2.4.1.

4.1 *Crude probabilities of death from cancer and other causes*

We started with the estimation of the pseudo-observations for CPr from cancer and other causes. Pseudo-observations for each cause (cervical cancer and other causes) were computed at 5 timepoints, which were decided based on the quantiles of the survival time distribution.

We modelled the CPr from cancer and other causes using three different models. All models were simple including time-dependent terms (i.e., indicator functions for the 4 last timepoints at which pseudo-observations were calculated) and two main variables, namely age at diagnosis and deprivation group. We used a linear term of age while deprivation was modelled as a categorical variable with 5 groups. Models differed with respect to link functions (cloglog, log, identity) allowing for different interpretations. The working covariance matrix was the same in all models where independence structure was applied. The estimated regression parameter estimates for each model can be seen in Table 3.

In the case of cloglog model, the reported $\hat{\beta}$ estimates correspond to log subdistribution hazard ratios associated with 1 unit change of a covariate X in the instantaneous rate of the occurrence of an event among those who are event-free or have experienced a competing event (i.e., the subdistribution hazard function). Following the reasoning in Section 2.4.1, we provided only a qualitative description of the results. Age coefficient is positive (0.452), which can be translated to an increase in subdistribution hazard and subsequently, in the probability of dying from cancer with the increase of age. Similarly, the regression parameter for age in the case of other causes was also positive (0.702), indicating an increase in the subdistribution hazard of other causes. Moreover, regardless of the COD, the most deprived people were associated with a bigger increase in CPr compared to the least deprived, with the only exception being those from deprivation group 2 in the cancer event. Lastly, we can say that for example people from deprivation group 4 who had a larger regression coefficient than those from deprivation 3, had a higher relative change in the incidence of cancer death (see Proof from Ref. (Austin and Fine, 2017)).

Although this interpretation was informative, the model with log link function provided additionally a quantitative interpretation expressed as relative risk. The effect of age was quantified as exp(0.33), meaning that a 10-year increase in age at diagnosis was associated with an increase in probability of death from cancer by 39% (95% CI:[37, 42]), for a given deprivation group at a given time-point. With respect to other causes, the regression parameter for the effect of a 10-year increase in age indicated an 1.95-fold (95% CI:[1.84, 2.08]) increase in the risk of dying from other causes. Regarding deprivation, by exponentiating the results shown in Table 3, we observed that the most deprived group (deprivation 5) had approximately 1.12 (95% CI:[1, 1.27]) times higher risk of dying from cancer compared to the least deprived group (deprivation 1) at a given time-point after adjusting for age at diagnosis. The corresponding effect on the other causes was 1.23 (95% CI:[0.97, 1.61]).

The identity link model has also the advantage of simple interpretation of the coefficient parameters, providing estimates of risk differences. Therefore, we observed that a 10-year increase in age at diagnosis was associated with an increase in the risk of cancer death (0.101, 95% CI:[0.095, 0.107]), for a given deprivation group at a given time-point (the corresponding estimates for other causes is 0.017 (95% CI:[0.015, 0.019])). Furthermore, we observed that for the most deprived group (deprivation 5) the risk difference related to death from cancer was estimated as 0.043 (95% CI:[0.02, 0.07]) compared to the least deprived group (deprivation 1) at a given time-point after adjusting for age. The analogous effect on the other causes was 0.008 (95% CI:[0.004, 0.012]). As we already mentioned in Section 2.4.1, one must be aware of inappropriate predictions when using this model which is true even here e.g. when trying to predict the probabilities for cancer for someone with an age below 38 years at the 1st time-point.

4.2 *Life years lost due to cancer and other causes*

The pseudo-observations for LYL from cancer or other causes were estimated within the time period 5 years. A GEE model with identity link function and independence working covariance matrix was applied with age at diagnosis and deprivation group as explanatory variables. According to the model estimates (see Table 4), a 10-year increase in age at diagnosis led to approximately 0.44 (95% CI:[0.42, 0.47]) additional years being lost due to cancer and 0.055 (95% CI:[0.051, 0.059]) due to other causes in the first 5 years. Moreover, people who were more deprived had an increased number of LYL compared to people who were less deprived in the first 5 years, with those in the most deprived group losing around 0.188 (95% CI:[0.08, 0.3]) additional years due to cancer compared to the least deprived.

5. Discussion

Alternative survival indicators such as CPr and LYL attributed to different causes can prove very useful when communicating survival statistics. That is especially true in the case where the event of interest is cancer whose complexity requires a multi-perspective approach. CPr and LYL are both defined in “real world” and quantify the impact of a covariate on a given event in the presence of other competing events thus, useful to inform about a patient’s prognosis, a treatment choice, or even the development of a control strategy (Charvat *and others*, 2013; Mariotto *and others*, 2014; Pohar Perme *and others*, 2016). The LYL indicator has the additional advantage of being expressed on a time scale, making it easier to communicate the results of analysis to a non-scientific audience (Belot *and others*, 2019). Although these indicators have been well defined and modelled in cause-specific setting, i.e., when the information on COD is available and reliable, a direct modelling of those measures in the relative survival setting was yet unavailable.

In this paper, we explored the use of pseudo-observations in modelling these alternative survival measures in relative survival setting with generalised linear models using the GEE method. This approach enables the user to choose between different link functions and various structures of working covariance matrix.

We evaluated the new approach using simulations and we showed that it performs well for both measures. Regarding CPr, assessment of different models through regression parameters showed good performance regardless of the choice of link function and whilst assuming a simple independence working covariance structure (Klein and Andersen, 2005; Pekár and Brabec, 2018). Regarding LYL, the simulation results displayed good performance for that indicator too, when applying an identity link function and an independence covariance matrix.

The application of the new method to cervical cancer data showed how the covariate effects on the indicators of interest can be derived and interpreted. The models used in the illustration were simple and model misspecification cannot be excluded yet, this study stresses on the interpretation rather than on model building strategies. One interesting further step would be to use goodness of fit tests as recently proposed (Pavlič *and others*, 2018), in order to assess the choice of link function and the functional form of continuous covariates.

In general, this approach offers a useful alternative, especially when considering how the interpretation simplifies when using a model for CPr with a log and identity link function (compared to one with a cloglog function). Although a cloglog link function would give similar interpretations to Fine & Gray model, we advocate the use of log link function with which $\exp(\beta)$ gives an estimate of relative risk (Overgaard *and others*, 2015), and of identity link function which would yield risk differences. This would avoid the pitfalls of interpreting subdistribution hazard ratios (Andersen *and others*, 2012; Austin and Fine, 2017) with the additional advantage of quantitative interpretation of covariate effects on the indicator of interest. However, one must be careful when choosing these link functions as to avoid predictions that go beyond the acceptable range (i.e.,[0,1] for probabilities and (0,+∞) for time).

Time-dependent and non-linear effects can also be easily introduced into the model (Klein and Andersen, 2005). However, inclusion of a time-dependent covariate needs careful consideration, mostly in terms of interpretation due to the fact that the CPr is not a functional of the sole intensity when (nondeterministic) time-dependent covariates are considered (Andersen *and others*, 2003). Knowing the future evolution of such covariates is therefore needed, yet this cannot be practically done when the observed COD is a competing event. Studies that deal with that include a landmarking approach using direct binomial modelling (Grand *and others*, 2018) or a synthesis of separate cause-specific hazard analyses (Beyersmann and Schumacher, 2008) etc., but more research in that direction will be needed in the context of pseudo-observation approach.

There are also other issues in our work which were not explored here, but which could be of possible interest. Firstly, until this point we presented a way to model the pseudo-observations separately for one cause at a time. An alternative choice would be to model them jointly and use a working covariance matrix that reflects the correlation between pseudo-observations of the same cause that would enable the joint estimation of parameters (Andersen, 2013). Secondly, the goal of this paper was to show the sensible behaviour of the method in practice. This was well confirmed with our simulations yet, more work is needed to to derive theoretically the asymptotic properties of the estimators. Thirdly, even though modelling pseudo-observations constitutes a simple and general approach that can simplify survival analysis, it is usually less efficient compared to other methods developed specifically for one indicator of interest. An additional consideration in this approach before applying it to any data, is the assumptions behind relative survival setting (Pavlič and Pohar Perme, 2018), violation of which might result in biased estimators of pseudo-observations and subsequently, an invalid analysis. Lastly, in this study we did not investigate the performance using different covariance matrix structures but we used the independence structure throughout as has been suggested by (Klein and Andersen, 2005). Impact of other structures on the results would be an interesting further methodological development.

In summary, our approach based on pseudo-observations in relative survival setting demonstrated nice frequentist properties on estimating the crude probabilities of death and the life years lost from different causes in realistic situations. These two indicators along with other frequently reported measures like net survival can improve the understanding of the nature and mechanism of competing events. Their computation in relative survival setting is quite important as routinely collected population-based data often suffer from unreliable or unavailable information of the COD. The advantage of the pseudo-observation approach to provide covariate effects directly affecting the indicators of interest in the relative survival setting, makes the method appealing to the user. However, one should be aware that this approach might be prone to a longer computational time (especially in the case of big datasets) compared to conventional methods. A guide that provides the code for applying the method in R-software can be found in Supplementary Material B.

6. Supplementary Material

The reader is referred to the Supplementary Materials A and B for technical appendices, R programs, and example output. Supplementary Material A can be found in Biostatistics online. Supplementary Material B, simulation code, and sample data are available at `https://github.com/pseudorel/supp_material.git`.

Acknowledgements

This research was supported by Cancer Research UK grant number C7923/A18525 and C7923/A20987. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of Cancer Research UK. Also, the authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P3-0154 and the project 'Years of life lost as a measure of disease burden', No. J3-1761).

Table 3. Regression parameter estimates (standard errors) for the direct modelling of the crude probabilities of death from cancer and other causes, as obtained with 3 models using pseudo-observations with link functions: *cloglog*, *identity* and *log*, and assuming an independence working covariance structure.

	cloglog		**log**		**identity**	
	cancer	*other causes*	*cancer*	*other causes*	*cancer*	*other causes*
(Intercept)	-2.313(0.072)	-5.835(0.149)	-2.178(0.055)	-5.743(0.14)	0.099(0.01)	0.002(0.001)
t= 969 days	0.626(0.028)	0.721(0.018)	0.477(0.023)	0.702(0.017)	0.098(0.004)	0.008(0)
t= 1826 days	0.845(0.031)	1.207(0.032)	0.612(0.025)	1.163(0.029)	0.14(0.004)	0.017(0.001)
t= 2132 days	0.881(0.032)	1.339(0.036)	0.633(0.026)	1.286(0.033)	0.147(0.004)	0.021(0.001)
t= 2487 days	0.927(0.033)	1.474(0.04)	0.659(0.026)	1.41(0.036)	0.156(0.005)	0.025(0.001)
Age†	0.452(0.014)	0.702(0.034)	0.33(0.009)	0.67(0.032)	0.101(0.003)	0.017(0.001)
Deprivation 2	-0.025(0.091)	0.158(0.147)	-0.016(0.067)	0.151(0.141)	-0.001(0.014)	0.002(0.002)
Deprivation 3	0.134(0.087)	0.099(0.137)	0.085(0.063)	0.094(0.131)	0.031(0.014)	0.003(0.002)
Deprivation 4	0.2(0.083)	0.135(0.126)	0.125(0.06)	0.125(0.121)	0.047(0.014)	0.005(0.002)
Deprivation 5	0.186(0.082)	0.223(0.129)	0.12(0.06)	0.203(0.124)	0.043(0.013)	0.008(0.002)

† (Age at diagnosis-47 (mean age in the dataset))/10

Table 4. Regression parameter estimates (standard errors) for the direct modelling of the number of life years lost due to cancer and due to other causes, as obtained with a model for pseudo-observations with *identity* link function and assuming independence working covariance structure.

	Cancer	*Other causes*
(Intercept)	0.841	0.051
Age†	0.443	0.055
Deprivation 2	0.002	0.005
Deprivation 3	0.144	0.011
Deprivation 4	0.216	0.017
Deprivation 5	0.188	0.025

† (Age at diagnosis-47 (mean age in the dataset))/10

References

Allemani, C, Matsuda, T, Di Carlo, V, Harewood, R, Matz, M, Nikšić, M, Bonaventure, A, Valkov, M, Johnson, CJ, Estève, J *and others*. (2018). Global surveillance of trends in cancer survival 2000–14 (concord-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet* **391**(10125), 1023–1075.

Andersen, PK. (2013). Decomposition of number of life years lost according to causes of death. *Statistics in medicine* **32**(30), 5278–5285.

Andersen, PK, Geskus, RB, de Witte, T and Putter, H. (2012). Competing risks in epidemiology: possibilities and pitfalls. *International journal of epidemiology* **41**(3), 861–870.

Andersen, PK, Hansen, MG and Klein, JP. (2004). Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis* **10**(4), 335–350.

Andersen, PK and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine* **31**(11-12), 1074–1088.

Andersen, PK, Klein, JP and Rosthøj, S. (2003). Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika* **90**(1), 15–27.

Andersen, PK and Pohar Perme, M. (2010). Pseudo-observations in survival analysis. *Statistical methods in medical research* **19**(1), 71–99.

Austin, PC and Fine, JP. (2017). Practical recommendations for reporting fine-gray model analyses for competing risk data. *Statistics in medicine* **36**(27), 4391–4400.

Belot, A, Abrahamowicz, M, Remontet, L and Giorgi, R. (2010). Flexible modeling of competing risks in survival analysis. *Statistics in Medicine* **29**(23), 2453–2468.

Belot, A, Ndiaye, A, Luque-Fernandez, MA, Kipourou, DK, Maringe, C, Rubio, FJ and Rachet, B. (2019). Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clinical epidemiology* **11**, 53.

Bender, Ralf, Augustin, Thomas and Blettner, Maria. (2005). Generating survival times to simulate cox proportional hazards models. *Statistics in medicine* **24**(11), 1713–1723.

Beyersmann, J, Latouche, A, Buchholz, A and Schumacher, M. (2009). Simulating competing risks data in survival analysis. *Statistics in medicine* **28**(6), 956–971.

Beyersmann, J and Schumacher, M. (2008). Time-dependent covariates in the proportional subdistribution hazards model for competing risks. *Biostatistics* **9**(4), 765–776.

Binder, N, Gerds, TA and Andersen, PK. (2014). Pseudo-observations for competing risks with covariate dependent censoring. *Lifetime data analysis* **20**(2), 303–315.

Charvat, H, Bossard, N, Daubisse, L, Binder, F, Belot, A and Remontet, L. (2013). Probabilities of dying from cancer and other causes in french cancer patients based on an unbiased estimator of net survival: A study of five common cancers. *Cancer epidemiology* **37**(6), 857–863.

Charvat, H, Remontet, L, Bossard, Ne, Roche, L, Dejardin, O, Rachet, B, Launoy, G and Belot, A. (2016). A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in medicine* **35**(18), 3066–3084.

Cronin, KA and 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.

Danieli, C, Remontet, L, Bossard, N, Roche, L and Belot, A. (2012). Estimating net survival: the importance of allowing for informative censoring. *Statistics in medicine* **31**(8), 775–786.

De Angelis, R, Sant, M, Coleman, MP, Francisci, S, Baili, P, Pierannunzio, D, Trama, A, Visser, O, Brenner, H, Ardanaz, E *and others*. (2014). Cancer survival in europe 1999–2007 by country and age: results of eurocare-5a population-based study. *The lancet oncology* **15**(1), 23–34.

Dignam, JJ and Zhang, M Qand Kocherginsky. (2012). The use and interpretation of competing risks regression models. *Clinical Cancer Research* **18**(8), 2301–2308.

Ederer, F. (1961). The relative survival rate: a statistical methodology. *NCI Monograph* **6**, 101–121.

Eloranta, S, Adolfsson, J, Lambert, PC, Stattin, P, Akre, O, Andersson, TML and 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.

Fine, JP and Gray, R J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association* **94**(446), 496–509.

Gerds, Thomas A, Scheike, Thomas H and Andersen, Per K. (2012). Absolute risk regression for competing risks: interpretation, link functions, and prediction. *Statistics in medicine* **31**(29), 3921–3930.

Geskus, Ronald B. (2015). *Data analysis with competing risks and intermediate states*. Chapman and Hall/CRC.

Grand, MK, de Witte, TJM and Putter, H. (2018). Dynamic prediction of cumulative incidence functions by direct binomial regression. *Biometrical Journal* **60**(4), 734–747.

Graw, F, Gerds, TA and Schumacher, M. (2009). On pseudo-values for regression analysis in competing risks models. *Lifetime Data Analysis* **15**(2), 241–255.

Haller, B and Ulm, K. (2014). Flexible simulation of competing risks data following prespecified subdistribution hazards. *Journal of Statistical Computation and Simulation* **84**(12), 2557–2576.

Hansen, SN, Andersen, PK and Parner, ET. (2014). Events per variable for risk differences and relative risks using pseudo-observations. *Lifetime data analysis* **20**(4), 584–598.

Hinchliffe, SR, Dickman, PW and Lambert, PC. (2012). Adjusting for the proportion of cancer deaths in the general population when using relative survival: a sensitivity analysis. *Cancer Epidemiology* **36**(2), 148–152.

Hjort, NL. (1992). On inference in parametric survival data models. *International Statistical Review/Revue Internationale de Statistique* **60**(3), 355–387.

Jacobsen, M and Martinussen, T. (2016). A note on the large sample properties of estimators based on generalized linear models for correlated pseudo-observations. *Scandinavian Journal of Statistics* **43**(3), 845–862.

Kipourou, DK, Charvat, H, Rachet, B and Belot, A. (2019). Estimation of the adjusted cause-specific cumulative probability using flexible regression models for the cause-specific hazards. *Statistics in medicine* **38**(20), 3896–3910.

Klein, JP. (2006). Modelling competing risks in cancer studies. *Statistics in medicine* **25**(6), 1015–1034.

Klein, JP and Andersen, PK. (2005). Regression modeling of competing risks data based on pseudovalues of the cumulative incidence function. *Biometrics* **61**(1), 223–229.

Lambert, PC, Dickman, PW, Nelson, CP and 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.

Lambert, PC, Wilkes, SR and Crowther, MJ. (2017). Flexible parametric modelling of the cause-specific cumulative incidence function. *Statistics in medicine* **36**(9), 1429–1446.

Liang, Kung-Yee and Zeger, Scott L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**(1), 13–22.

Mariotto, AB, Noone, A-M, Howlader, N, Cho, H, Keel, GE, Garshell, J, Woloshin, S and Schwartz, LM. (2014). Cancer survival: an overview of measures, uses, and interpretation. *Journal of the National Cancer Institute Monographs* **2014**(49), 145–186.

Moreno-Betancur, M and Latouche, A. (2013). Regression modeling of the cumulative incidence function with missing causes of failure using pseudo-values. *Statistics in medicine* **32**(18), 3206–3223.

Mozumder, SI, Rutherford, M and Lambert, P. (2018). Direct likelihood inference on the cause-specific cumulative incidence function: A flexible parametric regression modelling approach. *Statistics in medicine* **37**(1), 82–97.

Overgaard, M, Andersen, PK, Parner, ET *and others*. (2015). Regression analysis of censored data using pseudo-observations: An update. *Stata J* **15**(3), 809–21.

Overgaard, M, Parner, ET and Pedersen, J. (2018). Estimating the variance in a pseudo-observation scheme with competing risks. *Scandinavian Journal of Statistics* **45**(4), 923–940.

Overgaard, M, Parner, ET, Pedersen, J *and others*. (2017). Asymptotic theory of generalized estimating equations based on jack-knife pseudo-observations. *The Annals of Statistics* **45**(5), 1988–2015.

Pavlič, K, Martinussen, T and Andersen, PK. (2018). Goodness of fit tests for estimating equations based on pseudo-observations. *Lifetime data analysis* **25**, 1–17.

Pavlič, K and Pohar Perme, M. (2018). Using pseudo-observations for estimation in relative survival. *Biostatistics* **20**(3), 384–399.

Pekár, S and Brabec, M. (2018). Generalized estimating equations: A pragmatic and flexible approach to the marginal glm modelling of correlated data in the behavioural sciences. *Ethology* **124**(2), 86–93.

Pfeiffer, RM and Gail, MH. (2017). *Absolute Risk: Methods and Applications in Clinical Management and Public Health*. CRC Press.

Pohar Perme, M. (2018). Package "relsurv".

Pohar Perme, M, Estève, J and Rachet, B. (2016). Analysing population-based cancer survival-settling the controversies. *BMC cancer* **16**(1), 933.

Pohar Perme, M and Pavlic, K. (2018). Nonparametric relative survival analysis with the r package relsurv. *Journal of Statistical Software* **87**(1), 1–27.

Pohar Perme, M, Stare, J and Estève, J. (2012). On estimation in relative survival. *Biometrics* **68**(1), 113–120.

Rubio, FJ, Rachet, B, Giorgi, R, Maringe, C and Belot, A. (2019). On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables.

Satagopan, JM, Ben-Porat, L, Berwick, M, Robson, M, Kutler, D and Auerbach, AD. (2004). A note on competing risks in survival data analysis. *British journal of cancer* **91**(7), 1229–1235.

Talbäck, M and Dickman, PW. (2011). Estimating expected survival probabilities for relative survival analysis–exploring the impact of including cancer patient mortality in the calculations. *European journal of cancer* **47**(17), 2626–2632.

[*Received August 1, 2010; revised October 1, 2010; accepted for publication November 1, 2010*]