International Agency for Research on Cancer

World Health Organization

Université Grenoble Alpes

Rapport de stage

Net survival and cohort data

Survie nette et données de cohortes

Javier González Delgado

supervisé par

Vivian Viallon

Hadrien Charvat

Grégoire Rey

11 août 2020

Acknowledgements

I would like to thank my supervisor, Vivian, for his guidance, his time, and his kind and constant support. It has been a real privilege to learn from his scientific knowledge together with his exceptional human qualities. I would also like to thank my co-supervisors Hadrien Charvat and Grégoire Rey, whose active collaboration and help has been essential. A special thanks to Pietro Ferrari and my colleagues from IARC's NMB group, for their kindness and companionship.

Introduction

Cancer is currently the second most common cause of death in the world, reaching 9.8 million deaths in 2018 [1]. The recent medical progress made in sanitation, vaccination and antibiotic development, which had noticeably decreased mortality from infectious diseases, along with the improving mechanisms built to prevent cardiovascular diseases and the changes in demographic conditions and risk factors, had led cancer to be the first or second leading cause of early death (i.e. at 30-69 age years) in 134 out of 183 countries [1].

Cancer has a global but not an equal impact. Its patterns and trends in incidence and mortality vary notably across countries and specific cancer types. Those variations can be explained by both individual and structural factors, and lie mainly on differential exposures to proximal risk factors and on differences in access to health-care services. The profile of cancer type is one of the dominant disparities of cancer burden among countries, being closely related to the income level. Low-income countries have a higher incidence rate of infection-related cancers, whereas high-income countries present higher rates of other cancer types, such as prostate, breast, colorectal, thyroid and melanoma. A special case may be lung cancer, for which high incidence was firstly restrained to high-income countries, and has recently been recognised as a global scourge. Both in terms of indicidence and mortality, lung cancer is the leading cancer type with 2.1 million new cases and 1.8 million deaths in 2018 [1].

These variations on mortality for a given cancer type can be partly explained by the time and geographic changing mortality of the underlying general population, which is partially reflected on the cancer patients mortality. Thus, in order to correctly compare the efficiency of the health care system treatments among different countries or time periods, or to accurately measure the cancer burden among different populations, a mortality indicator that is independent of the influence of the general population mortality needs to be considered. That is why the concept of *net survival* has been introduced. Net survival corresponds to the survival that would be observed if the only possible cause of death was the considered disease (e.g. cancer). In other words, it corresponds to the survival observed in a hypothetical world where all the other causes of death would have been eliminated.

Quantifying cancer net survival is an important but challenging task. Two main methods of survival analysis are generally used: cancer-specific survival and relative survival. The former relies on the information on causes of deaths made available by death registries, while the latter uses deaths from any cause and compares the observed survival in cancer patients with the one of the general population. Some authors have suggested that relative survival is the most, and probably the only, adequate measure to use in cancer survival studies [2–5]. The basis of their argument is that misclassification of cancer deaths may occur and therefore biased estimates may appear when using cancer-specific survival. However, there has not been a clear evidence that using causes of deaths is not an appropriate way to measure net survival when high quality data is available, or that relative survival estimation is always a more accurate solution.

This project will aim to provide both relative and cause-specific estimates of European lung and colorectal cancer net survival, using data from the whole EPIC cohort (European Prospective Investigation into Cancer and Nutrition). Having access to this prospective study entails an a priori leap of the main sources of bias of both relative and cause-specific estimation methods, as we will further discuss, as well as the possibility to work with high quality survival data. Both types of estimates will be computed in order to assess the differencies which would eventually appear, the quality of the estimates, the pertinence of causes of death in order to estimate cancer survival, and the possible effect of covariates as tobacco smoking. To do so, standard (net) survival analysis methods will be implemented, as well as modern statistical tools such as high flexible parametric models.

All the developped work and research has been done as a part of a Master 2 internship at the International Agency for Research on Cancer (IARC). The International Agency for Research on Cancer is the specialized cancer agency of the World Health Organization (WHO). Its objective is to promote international collaboration in cancer research, bringing together skills in epidemiology, laboratory sciences and biostatistics to identify the causes of cancer so that preventive measures may be adopted and the burden of disease and associated suffering reduced. A significant feature of IARC is its expertise in coordinating research across countries and organizations, for which its independent role as an international organization is essential.

More particularly, the project has taken place at the Nutritional Methodology and Biostatistics Group (NMB) of the Nutrition and Metabolism Section (NME). The goal of the NME section, led by Dr Marc Gunter, is to provide robust scientific evidence on the role of nutrition, obesity, and metabolic dysfunction in cancer development that can translate to both clinical and population-level interventions and to public health policy. The NMB group, led by Dr Pietro Ferrari, fosters the methodological work that is crucial to integrate and optimize the use of these resources for studies of cancer prevention. This project was directly supervised by Dr Vivian Viallon, chair of the IARC Statistical Working Group. Dr Viallon focuses on methodological developments to answer epidemiological questions, and his main research topics involve the study of obesity and cancer risk, the impact of comorbidity on cancer diagnosis and survival and the cancer (net) survival analysis in competing risks or multistate settings. In the context of the latter, Dr Viallon works together with Dr Grégoire Rey from the INSERM's Centre d'Épidémiologie sur les causes médicales de décès (CépiDc) and Dr Hadrien Charvat from IARC's Section of Cancer Surveillance (CSU), who have co-supervised this project.

In this report, we begin with a more technical introduction on net survival and the presentation of methods available for its estimation. In particular, we will discuss the possible sources of bias or error attached to these different methods, and we will describe in detail the estimators and models that we have implemented along this project. After briefly presenting the EPIC cohort and its main characteristics, we will illustrate their application for the estimation of the net survival among lung cancer patients in the EPIC cohort. Relative and cause-specific methods will be first introduced separately, and the results will be jointly discussed.

Contents

1 Net survival: concept and estimation

1.1 Survival analysis in the competing risks setting

Survival analysis is the branch of statistics which analyses the time until some event of interest occurs. Depending on the field of study, this event may be the failure time of an electronic system, the time a person leaves a job, or the death of a patient diagnosed with a specific disease. The main difficulty when analysing survival data is often the *censoring* phenomenon, which leads to a partial observation of survival times. This may happen, for instance, when an invidual leaves the study without experiencing the event of interest. However, even if the information is partial, specific statistical tools and methods can provide unbiased estimates and accurate predictions.

Standard survival analysis can be considered from a more general setting where multiple target states are possible. This is what is called a *competing risks* analysis, where the composite endpoint would be distinguished by investigating the time to event and the event type, and only transitions between a common initial state and the competing risks states would be considered. A clinical oncology example is progression-free survival, which is the time until disease progression or death, whichever occurs first. If we also wanted to consider death after progression, or a possible recovery of patients, the more general setting of *multistate models* would be required.

Classical survival analysis corresponds to the simplest multistate model, displayed in Figure 1, where an individual is in the initial state 0 (alive) at time origin, and at some later random time T, the individual moves to the absorbing state 1 (dead).

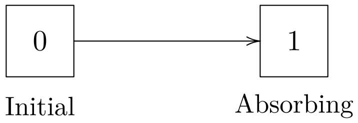


Figure 1 – The most simple survival multistate model.

We are interested in the event time T, which is a continuous random variable often called *survival time* or *failure time.* The statistical analysis of T is usually based on the hazard rate $\lambda(t)$ attached to the distribution of T and defined as the limit

$$\lambda(t) = \lim_{dt \to 0} \frac{1}{dt} \mathrm{P}(t \leq T < t + dt \,|\, T \geq t). \tag{1}$$

The hazard function λ determines the distribution of T and thus its usual characterisations:

$$F(t) = 1 - S(t) = \mathrm{P}(T \leq t) = 1 - \exp\left(-\int_0^t \lambda(u)du\right), \qquad f(t) = \lambda(t)\,S(t), \tag{2}$$

where F is the cumulative distribution function of T, f the density function of T and S the survival function of T. The reason why survival analysis is hazard-based is that hazard function

λ remains undisturbed by censoring [6]. In real clinical studies, data analysis is frequently performed before or without knowing all end-point times (e.g. many patients surviving after the study closure, individuals dropped out of the study beacause they move to a different place, etc.), which leads to incomplete observations (*right censoring*). The usual estimation tools, such as the empirical survival function, provide biased estimates, but it is easy to prove that, if we introduce an independent (of the event time) censoring time C, and thus we would be observing the pair $(\min(T,C)\,,\,\mathrm{I}_{T\leq C})$, hazard function λ satisfies:

$$\lambda(t)\,dt = \mathrm{P}(t \leq T < t+dt \,|\, T \geq t) = \mathrm{P}(t \leq T < t+dt\,,\, T \leq C \mid \min(T,C) \geq t) \quad \forall\, t \geq 0, \tag{3}$$

so censoring has not disturbed the hazard. This fact has a number of important consequences. First, the estimation of the cumulative hazard

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{4}$$

is closely connected to counting processes and martingales. A counting process merely counts the number of observed events with the passage of time. Martingale theory provides estimating equations and small and large sample properties of estimators. [7, 8]. This connection enables the analysis of event time data in settings that go beyond the right-censored and single-event type situation. For example, the hazard-based approach is also able to deal with *left-truncated* data, where patients have delayed their study entry times.

As mentionned above, this hazard-based framework may be generalized to a competing risks setting, under which net survival theory may be built. The two-state survival model of Figure 1 can be now extended to competing risks by introducing several competing target states representing each one of the possible event types. Occurence of a competing event is modelled by a transition towards the corresponding competing event state. Such a model is illustrated in Figure 2a with a finite number M of competing risks. Figure 2b depicts the corresponding model in the particular case of two competing risks.

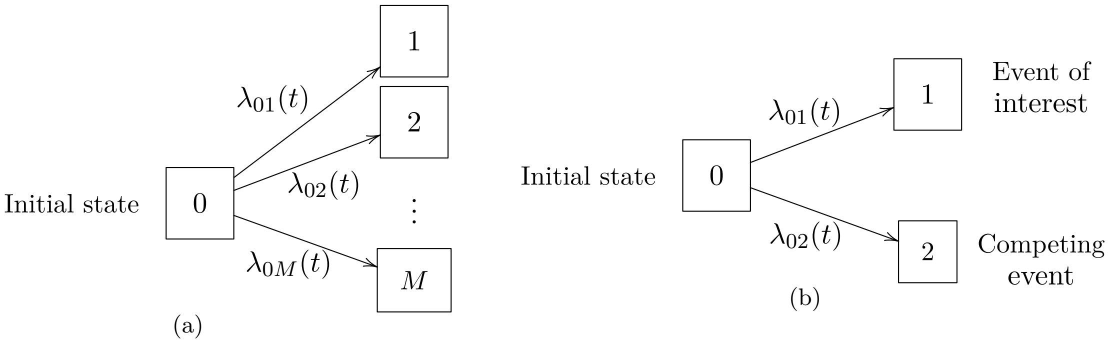


Figure 2 – Competing risks model with M (a) and 2 (b) event-specific hazards.

We may now define X_t as the position occupied by the process at time $t \geq 0$ and, as before, the event time T as the earliest time at which the individual is not in the intial state 0 anymore: $T = \inf\{t : X_t \neq 0\,,\, t \geq 0\}$. In addition to the survival time, competing risks data involves a second component, the event type, denoted by X_T. Our data consist now of the pair $(T\,,\, X_T)$ with $X_T \in \{1, \ldots, M\}$. As displayed in Figure 2, there is now one event-specific hazard per competing event, defined as

$$\lambda_{0j}(t) = \lim_{dt \to 0} \frac{1}{dt} \mathrm{P}(t \leq T < t + dt\,,\, X_T = j \,|\, T \geq t), \quad j = 1, \ldots, M. \tag{5}$$

The interpretation of (5) is that $\lambda_{0j}(t)dt$ is the probability that a type j event takes place in the infinitesimal time interval $[t\,,\, t + dt)$, conditional on the fact that no event (of any type) has occured before t. As in the standard survival analysis framework, left-truncation and right-censoring mechanisms do not disturb the event-specific hazards λ_{0j} if they are independent of the event time. The event-specific cumulative hazard and survival function can be defined paralleling (2) and (4):

$$\Lambda_{0j}(t) = \int_0^t \lambda_{0j}(u)du \qquad S_{0j}(t) = \exp(-\Lambda_{0j}(t)). \tag{6}$$

In order to estimate Λ_{0j} through a counting process one should code type $j' \neq j$ events as censoring events and just type j events as actual events: occurrence of type j' events take place as independent right-censoring in respect of j events. Consequently, the estimation of Λ_{0j} can be performed by removing the type j' events from the risk set. A more rigorous discussion of this issue can be found in [7].

For our net survival analysis we will be placed both in Figure 1 and Figure 2b settings, depending on which approach we will be considering. Net survival and its connection to the competing risks settings are presented in the next section. We will further describe common approaches for the estimation of net survival, and briefly discuss their systematic bias and possible sources of error.

1.2 Relative survival vs cancer-specific survival

Cancer survival (i.e. the 'overall' survival of cancer patients) is a key tool when analysing, understanding and quantifying cancer burden. In order to identify inequalities among countries, test the efficiency of cancer treatments or the effect of comorbidities on the disease, assessing the behaviour of cancer survival is crucial. However, not all cancer patients die because of their cancer, so that cancer survival is not only determined by the cancer itself (i.e. by the site and stage of tumor, the diagnosis date, etc.) but also by a great amount of genetic, demographic, and lifestyle factors, which rely on the time-varying general population mortality. Consequently, cancer survival is not an ideal criterion to properly quantify the mortality due to the cancer itself.

This is why the concept of *net survival* has been introduced, which is defined as the survival that would be observed if cancer was the only possible cause of death. This measure is clinically meaningless, as it computes the survival of a hypothetical population only exposed to cancer, but it is the only indicator of survival that can be used to accurately measure cancer burden independently of other causes of death. If we denote by λ_{Cancer} the hazard function associated to this hypothetical population, net survival is therefore defined as:

$$S_{\text{Net}}(t) = S_{\text{Cancer}}(t) = \exp\left(-\int_0^t \lambda_{\text{Cancer}}(u)du\right). \tag{7}$$

As we mentionned above, two main methods have been introduced to calculate cancer net survival: cancer-specific survival and relative survival. Both methods focus on mortality among cancer patients. Cancer-specific survival uses cancer-specific deaths as the end-point of interest, and patients who die from other causes are considered to be censored. Relative survivala uses death from any cause as the end-point of interest, and compares the observed survival with that which would have been expected it the cancer patients had had the same mortality rates as the general population. A number of authors have argued that relative survival is the most, and possibly the only, appropriate measure to use in population-based cancer survival studies [2–5]. The basis of this argument is that misclassification is expected in the causes of death, resulting in biased estimates of cancer-specific survival. However, relative survival is also susceptible to biased and unaccurate estimates as it uses the general population, which may not always be comparable to the cancer cohort (this will be discussed in detail in section 1.2.2). We provide now an overview of the two methods and their assumptions, and what type of systematic error may affect each one.

1.2.1 Cancer-specific survival

Cancer-specific (or cause-specific) survival analysis consists in focusing on deaths identified as being due to a specified cancer as the outcome of interest, and its implementation is based on the information of death certificates provided by physicians. This directly erases the influence of general population on cancer mortality, and place us on the competing risks setting of Figure 2b which, for the cancer-specific framework, may be reconsidered as:

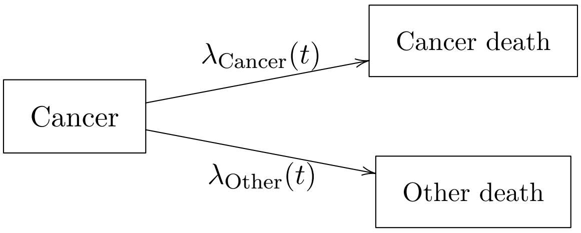


Figure 3 – Cancer-specific multistate model.

Follow-up starts on the date of cancer diagnosis, and continues until death, loss-to-follow-up or the end of the study period, whichever occurs first. Patients who die from a cause other that the cancer under study are censored. The survival for a given time period can be therefore directly calculated:

$$S_{\text{Cancer}}(t;x,a) = \exp\biggl(-\int_0^t \lambda_{\text{Cancer}}(u;x,a)du\biggr), \tag{8}$$

where x is a set of covariates and a the age at diagnosis. Therefore, information on causes of death place ourselves in a setting where net survival estimation is direct, as (8) corresponds to net survival definition (7). However, special attention must be made when using standard methods as Kaplan-Meier [9, 10] that require non-informative censoring, as this condition may not be verified when censoring for other causes deaths [11, 12].

The most important potential source of bias in relation to estimating cancer-specific survival is causes of death misclassification [4]. This may happen if there is a lack of sensitivity or specificity on cancer death classification, which can be minimized with high-quality data, or because of the intrinsic difficulty of causes of death determination, which is a conceptual and more delicate issue. Consider for instance a patient on hormone treatment for breast cancer who dies of a pulmonary embolism or a patient having a successful curative lobectomy for lung cancer but diying two years later from pneumonia. In these cases the cancer is likely to have contributed to the death to some extent, but it is impossible to accurately ascribe such individuals deaths as being wholly cancer specific or not. To deal with this, each death is classified by identifying a single, *underlying* cause, which is defined as the disease or injury that initiated the train of events leading directly to death. In order to standardize the procedure of single-cause attribution when more than one condition contributes to death, World Health Organization has stipulated a methodology to properly determinate the underlying cause of death from the information of death certificates [13].

It should also be mentionned that some cancer registries might not have access to data on specific cause of death, and therefore cancer-specific survival cannot be computed. Nonetheless even registries that do have such data are still dependent on the quality of the causes of death. If causes of death are unavailable or have a dubious quality, relative survival framework provides an alternative way to get rid of the influence of general population which does not require the causes of death information.

1.2.2 Relative survival

Relative survival framework is based on the setting of unavailable or unreliable causes of deaths, where deaths due to any cause are the ones considered in the cancer patients cohort. We are now placed in Figure 1 model, which can be reconsidered as:

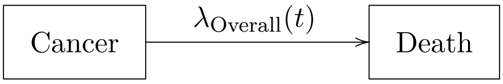


Figure 4 – Relative survival multistate model.

The advantage of this framework is that a net cancer survival can be estimated whithout the information of the causes of death in the cohort. This approach is based on the decomposition of the observed (overall) hazard of deaths among cancer patients into the hazard of deaths due to the disease and that of deaths due to other causes:

$$\lambda_{\text{Overall}}(t; x, z, a) = \lambda_{\text{Cancer}}(t; x, a) + \lambda_{\text{Other}}(a + t; z), \tag{9}$$

where t is the time since cancer diagnosis, a is the age at diagnosis and x and z are sets of covariates explaining respectively the cancer-related and the general population mortality. This decomposition establish a connection between Figure (3) and Figure (4) settings, allowing λ_{Cancer} estimation. The observed hazard λ_{Overall} is directly estimated from the cancer registry and the expected hazard λ_{Other} is estimated from an external population, whose general mortality (i.e. all-causes mortality excluding the cancer-specific one) has to be comparable to the cancer cohort one, which is quite a strong assumption. This is often worked out using general population life tables. Having both observed and expected hazard estimates allows λ_{Cancer} estimation and thus net survival calculation:

$$S_{\text{Cancer}}(t; x, z, a) = \exp\left(-\int_0^t (\lambda_{\text{Overall}}(u; x, z, a) - \lambda_{\text{Other}}(u + a; z))du \right). \tag{10}$$

The most important source of bias specific to relative survival analysis is the potential lack of comparability between the cancer cohort and the external population. The assumption of comparability will no longer stand if a factor that influences mortality from other causes is differently distributed between the cancer and external group. For instance, patients with smoking-related cancers will have a remarkable higher tobacco exposure compared with the general population, so their risk of death from other tobacco-related conditions will be considerably greater [14]. Other lifestyle related or demographic variables may also cause non-comparability, such as obesity, ethnicity or socio-economic position. As those covariates are generally not available in general population life tables nor in cancer registries, a potential bias in net survival estimation is likely to appear when comparability is not guaranteed, and an alternative estimation of an adapted λ_{Other} should be considered. For further technical details of non-comparability see [12].

1.3 Mixed-Effect Excess Hazard Regression Models (mexhaz)

Several methods have been implemented in the framework of this project to estimate net survival. We will focus on parametric estimation via a flexible regression model developped by Charvat et al. [15] that we considered is the more pertinent, adaptive and easy to implement for our analysis. Other estimation methods were also assessed and implemented in our study, notably the Poisson regression model and the Pohar-Perme non-parametric estimator. A fully detailed description of these methods, and an analysis of when and why they may be computed to estimate net survival, can be found in the Appendix A 'Further estimation methods of net survival'.

Modelling the excess hazard λ_{Cancer} is one of the approaches developed to estimate net survival [16–21], and it has been shown to provide unbiased estimates of net survival as long as time dependent and non-linear effects of relevant covariates are modelled [11]. We will present in this section the main method used in the framework of this project to model net hazard and survival fulfilling the just mentionned requirements.

Charvat et al. [15] developped an approach to fit a (net) hazard regression model allowing the flexible and non-proportional modelling of covariates and including a random effect at the cluster (e.g. country or other geographical area) level. This random effect is normally distributed, and it can be included to model the unobserved heterogeneity between clusters, where individuals show a shared frailty towards the disease (which explains the possible correlation of their survival times).

The model was primarily developped to estimate relative survival, although it can also be used to estimate cause-specific and overall survival. In all cases, the corresponding hazard is modelled as a function of time and a set of covariates depending on a vector of parameters β, and the baseline hazard is modelled as a B-spline or a restricted cubic spline[1]. The covariates' time-dependent effects are modelled as interaction terms between the covariates and the time scale, whose functional form will therefore determine the the time-dependent effect one. Non-linear effects of covariates can also be included in the model. The general expression for the (net) hazard is

$$\lambda(t;x)\exp(w) = \exp\left[\alpha_0 + \sum_{i=1}^{P+Q} \beta_i\, f_i(x_i) = + \sum_{j=1}^{R}\left(\gamma_{j0} + \sum_{k=P+1}^{P+Q} \gamma_{jk}\, f_k(x_k)\right) F_j(t)\right]\exp(w) \quad (11)$$

where:

· w is the random effect at the cluster level,

[1] Spline regression is an efficient alternative to polynomial regression based on piecewise polynomial functions which provides accurate evaluations and keeps the computational advantages of linearity and the flexibility of local polynomials. For further technical details see [33] or [34].

- $F_j(t)$, for $j = 1, \ldots, R$, are the basis functions of time used to describe the baseline hazard and the time dependent effects of covariates,
- α_0 is the intercept of the model,
- β_i, for $i = 1, \ldots, P$, are the coefficients corresponding to the covariates modelled with a proportional effect,
- β_i, for $i = P+1, \ldots, P+Q$, are the coefficients corresponding to the non-time dependent part of the effect of the covariates modelled with a time-dependent effect,
- γ_{j0}, for $j = 1, \ldots, R$, are the coefficients corresponding to the spline modelling the logarithm of the baseline hazard,
- γ_{jk}, for $j = 1, \ldots, R$ and $k = P+1, \ldots, P+Q$, are the coefficients corresponding to the modelling of the time-dependent effect of the covariates (obtained by considering interactions terms with the function used to model the baseline hazard),
- f_i, for $i = 1, \ldots, P+Q$ are the functions defining the non-linear effects of covariates, determined by a spline function. Linear effects can be fixed by setting $f_i = \mathrm{I}$.

Let's see how this model looks like in a simple framework. Consider a baseline hazard defined by a quadratic B-spline with two knots (i.e. requiring four basis functions, named here $BS_1, \ldots, BS_4$, in addition to the intercept), a first covariate x_1 modelled with a proportional (i.e. constant in time) effect, and a second covariate x_2 modelled with a time-dependent effect. If we consider a fixed effects model (i.e. with no random effect), equation (11) becomes:

$$\lambda(t; x_1, x_2) = \exp\left(\alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \sum_{j=1}^{4} \gamma_{j0} BS_j(t) + x_2 \sum_{j=1}^{4} \gamma_{j2} BS_j(t)\right) = \tag{12}$$

$$= \underbrace{\boxed{\exp\left(\alpha_0 + \sum_{j=1}^{4} \gamma_{j0} BS_j(t)\right)}}_{\lambda_0(t)} \exp\left(\beta_1 x_1 + \underbrace{\boxed{\left(\beta_2 + \sum_{j=1}^{4} \gamma_{j2} BS_j(t)\right)}}_{f(t)} x_2\right), \tag{13}$$

and thus the hazard can be expressed in the clearer form:

$$\lambda(t; x_1, x_2) = \lambda_0(t) \exp(\beta_1 x_1 + f(t) x_2), \tag{14}$$

with the restriction that f is based on the same basis function of time as the ones used to model the logarithm of the baseline hazard $\lambda_0(t)$.

The proposed estimation method of the model parameters is based on a likelihood maximisation. To do so, if we denote by t_{ij} the survival time and δ_{ij} the event indicator for the individual

$j = 1, \ldots, n_i$ from cluster $i = 1, \ldots, C$, the net hazard model is defined as

$$\lambda_{Overall}(t; x_{ij}, z_{ij}, w_i) = \lambda_{Cancer}(t; x_{ij}) \exp(w_i) + \lambda_{Other}(a + t; z_{ij}), \tag{15}$$

where w_i is the random effect at cluster level. Paralleling the covariates notation of section 1.2.2, the likelihood for a single observation (t_{ij}, δ_{ij}) from cluster i conditional on the value of the random effect is then

$$L_{ij}(\beta \,|\, w_i) = (\lambda_{Cancer}(t_{ij}; x_{ij}, w_i) + \lambda_{Other}(a + t_{ij}; z_{ij}))^{\delta_{ij}}\, S(t_{ij}; x_{ij}, z_{ij}, w_i), \tag{16}$$

where all the parameters have been grouped in a single vector β and

$$S(t_{ij}; x_{ij}, z_{ij}, w_i) = \exp(-\Lambda_{Cancer}(t_{ij}; x_{ij}, w_i) - \Lambda_{Other}(a + t_{ij}; z_{ij})). \tag{17}$$

In practice, the last term of the exponential in (17) can be removed from the estimation procedure as it does not depend on the parameters to be estimated. The marginal likelihood for cluster i is obtained by integrating the conditional likelihood for cluster i over the (normal) distribution of the random effect:

$$L_i(\beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \prod_{j=1}^{n_i} L_{ij}(\beta, w_i) \exp\left(\frac{w^2}{2\sigma^2}\right) dw, \tag{18}$$

and then the model parameters (β, σ) can be estimated by maximising the full log-likelihood:

$$\log(L(\beta, \sigma)) = \sum_{i=1}^{C} \log(L_i(\beta, \sigma)). \tag{19}$$

In order to compute overall or cause-specific hazard, λ_{Other} has to be set to zero. When estimating net hazard, $\lambda_{\text{Other}}(a + t_{ij})$ values for each individual and cluster are considered to be given, with no standard deviation, and they can be obtained directly from life tables, or from the modelling of λ_{Other} on the non-cancer population. The fact that $\lambda_{\text{Other}}(a + t_{ij})$ values are considered by the model as perfect estimates has to be taken into account when interpreting λ_{Cancer} confidence intervals, which will be showed narrower than they really are.

Cumulative hazard integrals can not be calculated analytically, and therefore a numerical procedure has to be carried out during the optimization process and also in order to compute survival estimates. Charvat et al. also developed the `R` package `mexhaz` where the introduced method to model (net) hazard and survival is implemented [22]. Adaptive Gauss-Hermite quadrature is the chosen method to compute integrals, and maximisation is performed using the `R` function `nlm` based on the Dennis-Schanabel non-linear unconstrained minimiser [23]. For further details on `mexhaz` model and its implementation in `R` see [15] and [22].

2 EPIC cohort

The European Prospective Investigation into Cancer and Nutrition (EPIC) study [24] is one of the largest cohort studies in the world, with more than half a million (521,000) participants recruited across 23 European centers and followed for almost 15 years. EPIC was designed to investigate the relationships between diet, nutritional status, lifestyle and enviromental factors, and the incidence of cancer and other chronic diseases. The EPIC study is jointly coordinated by Professor Elio Riboli, Director of the School of Public Health at Imperial College of London, United Kingdom, and Dr Marc Gunter and Dr Paul Brennan at IARC.

From the recruitment of study participants in 1992–1999 until 2015, the cohort accumulated more than 8 million person-years. More than 67,000 EPIC participants were diagnosed with cancer, including about 16700 cases of breast cancer, 4,600 of lung cancer, 7,100 of colorectal cancer, and 7,500 of prostate cancer. Also, 58,000 deaths were reported.

Therefore, the EPIC cohort provides access to very rich data to study causes of cancer, and cancer survival. The database used in this project contains follow-up information of 501,665 individuals, of which 58,318 were diagnosed with cancer, and up to 75 variables describing dietary exposure, lifestyle factors, anthropometry and biological parameters. Among the strong points of EPIC data is the access to high quality cause of death information, to which the colaboration of Grégoire Rey (CépiDc-INSERM) has been essential. This allows the implementation of both relative and cause-specific survival estimation methods, and the analysis of the effect of a great amount of covariates on cancer survival or incidence. Having both accurate cause of death data and covariates information of the individuals is not common among cancer registries and it makes EPIC a precious cohort for a very large variety of studies.

3 Net survival estimation in the EPIC cohort

The main objective of this project was to compare estimates of net survival produced by the cause-specific and relative cancer survival approaches in the EPIC cohort. In this section, we will introduce in detail how both approaches have been implemented, and particularly how EPIC data allowed us to deal with non comparability between the general population and the cohort when computing relative survival. We start with the presentation of the relative survival estimation in the EPIC cohort. We first implement the relative survival method using available life tables for each country represented in EPIC to estimate λ_{Other}. However, as most cohorts, the EPIC population is not comparable with the general population (it suffers from the *healthy-cohort* effect). We therefore used data from the EPIC cohort to estimate λ_{Other}. As a by-product, the comparison between this estimate and the one available in life tables allowed a precise quantification of the healthy-cohort effect in EPIC in terms of mortality. Then, we give details on the implementation of the cause-specific survival approach and, finally, we present and discuss some results produced by the two approaches.

3.1 Relative survival estimation

The relative survival framework uses all-cause death among cancer patients as the endpoint of interest. The cancer cohort (i.e. the individuals diagnosed with the cancer of interest) are used to compute overall hazard $\lambda_{Overall}$, and the excess hazard λ_{Other} is estimated on an external (general) population, which must be comparable in order to get unbiased estimates. That means that the mortality due to other causes has to be the same for the individuals of the general population and for those of the cancer cohort with the same demographic characteristics (i.e. birth year, sex, country, etc.). On one hand, what we do observe on EPIC is the process:

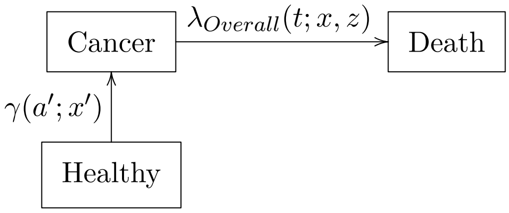


Figure 5 – Survival multistate model observed on EPIC's relative survival framework.

where t is the time since diagnosis, a' is the age of healthy individuals, γ is the hazard associated to the time to cancer diagnosis of healthy individuals and x, x' and z are three sets of covariates. On the other hand, what we can relate from general population is:

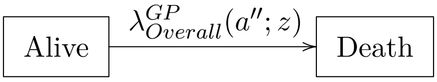


Figure 6 – Survival multistate model observed on general population.

where a'' is the age of the general population individuals. First, the proportion of individuals diagnosed with the cancer of interest among general population is negligible, so we can fairly estimate the general population other cause mortality by the all causes one: $\lambda^{GP}_{Other} \approx \lambda^{GP}_{Overall}$. Then, if we can assume

$$\lambda_{Other}(u; z) \approx \lambda^{GP}_{Other}(u; z) \qquad \forall\, u \geq 0, \tag{20}$$

we may estimate net survival from the decomposition:

$$\lambda_{Overall}(t; x, z) = \lambda_{Cancer}(t; x) + \lambda^{GP}_{Overall}(a + t; z), \tag{21}$$

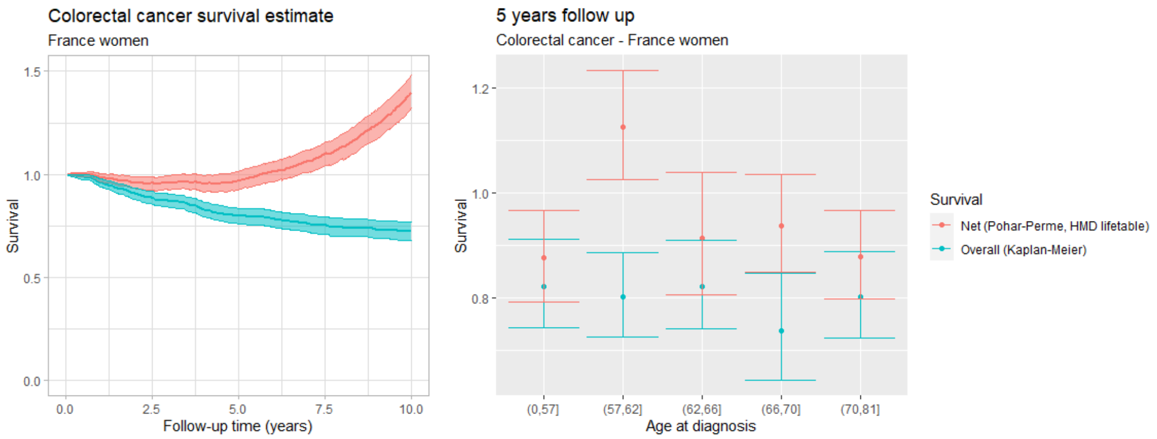
where a denotes the age at diagnosis. However, the comparability condition (20) is not often satisfied on cancer registries, as cancer patients clearly do not constitute a random sample of the general population, but are selected according to medical requisites. Our first objective was to explicitly assess this selection effect, and estimate net survival using general population lifetables to compute the excess hazard.

3.1.1 General population expected hazard

We aimed here to provide non-parametric relative survival estimates using the Pohar-Perme estimator [11] (see Appendix A.0.1). In order to compute excess hazard, we used the Human Mortality Database (HMD) life tables [25], which provide general population mortality rates per calendar year, age, country and sex for 41 countries or areas, and the analogous life tables from the CONCORD programme [26]. One of the advantages of using Pohar-Perme estimator is that its implementation in R requires a general population life table to be passed as an argument, which notably facilitates coding in this first case.

In order to assess the selection effect while estimating EPIC relative survival by computing excess hazard from general population life tables, we implemented Pohar-Perme estimator for individuals diagnosed with colorectal cancer, stratified by age at diagnosis. The results for the French women cohort using HMD life tables are illustrated in Figures 7a and 7b, where the Kaplan-Meier estimate of overall survival is also included to facilitate comparison.

Figure 7a shows the overall and relative survival curves for the EPIC's French women cohort. The selection effect is clearly identifiable after the net survival trend, which is monotonically increasing from 5 years of follow-up and reaches values higher than 1, which is not coherent with survival definition. Same unconsistent behaviour can be found when stratifying by age at diagnosis groups (Figure 7b). This incoherence is stronger or weaker depending on the selected country, sex and cancer type. Among colorectal cancer, French women show the most drastic selection effect, while Norwegian women show the littlest. However, we constate this trend among all countries and sexes, and thus general population mortality is not comparable to the other causes mortality of EPIC cancer patients. In particular, after the revealed increasing trends, general population hazard overestimates EPIC's λ_{Other}, as EPIC individuals diagnosed with cancer are dying less of any cause than the general population does. Therefore, *Other* is the most powerful absorbing state and *Cancer* shows an (irreal) extremly high specific survival.



(a) Relative and overall survival curves. (b) Relative and overall survival estimates at 5 years of follow-up.

Figure 7 – Colorectal cancer relative and overall survival estimates for EPIC French women.

3.1.2 Relative survival estimation with EPIC based expected hazard

After confirming the non comparability of general population and EPIC cancer cohort, an appropriate expected hazard must be built in order to properly estimate relative survival. As mentionned above, selection effect can be avoided as our database contains follow-up information of 397,756 individuals who have not been diagnosed with any type of cancer. We therefore have access to a comparable population in terms of other causes mortality, which is the subset of EPIC cohort who have not been diagnosed with the cancer of interest. The plan is now very clear, and relies on adapting Figure 5 to the current setting:

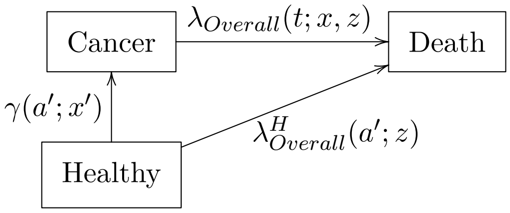


Figure 8 – Survival multistate model observed on EPIC's relative survival framework.

where we are considering as 'Healthy' all the individuals who are not diagnosed with the specific cancer of interest, and denoting by $\lambda^{H}_{Overall}$ their death hazard. If we place ourselves at the initial state 'Healthy', and consider the two possible end-points 'Cancer' and 'Death', we can adress $\lambda^{H}_{Overall}$ estimation as a competing risks problem: age of death of 'healthy' EPIC individuals is the outcome of interest, and individuals who develop the cancer of interest are censored at the age of their cancer diagnosis. If $\lambda^{H}_{Overall}(u; z) \approx \lambda_{Other}(u; z)$ for all $u \geq 0$, then we can apply the relative survival approach which relies on the decomposition:

$$\lambda_{Overall}(t; x, z) = \lambda_{Cancer}(t; x) + \lambda^{H}_{Overall}(a + t; z). \tag{22}$$

Both the expected and net hazards, $\lambda^{H}_{\text{Overall}}(a+t; z)$ and $\lambda_{\text{Cancer}}(t; x)$ can be estimated under the flexible hazard regression model (`mexhaz`) developped by Charvat et al. [15]. Expected hazard will be modelled using birth year, sex, and country or center as covariates. Baseline hazard and the effect of birth year will be modelled as the exponential of a 3 degree B-spline with two knots at the 1/3 and 2/3 quantiles of the event ages and the birth years, respectively. All the two-way interactions and time-dependent effects will be included in the initial model, and we will finally only retain the significant ones according to *backward* variable selection procedure based on the AIC criterion. The formal expression of expected hazard is therefore:

$$\lambda_{\text{Other}}(a'; z) = \underset{\text{B-spline}}{\boxed{\lambda^{0}_{\text{Other}}(a')}} \exp \sum_{i=1}^{p=3} \left(\beta_i(a') f_i(z_i) + \frac{1}{2} \sum_{j=1, j \neq i}^{p=3} \beta_{j,i}(a') f_j(z_j) f_i(z_i) \right), \tag{23}$$

where a' is the age of the individuals who have not been diagnosed with the cancer of interest, z_1 is their sex (male or female), z_2 their country or center and z_3 their birth year. As sex and country are modelled with a linear effect, $f_1 = f_2 = \mathrm{I}$, whereas f_3 corresponds to the B-spline used to model the effect of birth year. The equivalent Poisson regression model (see Appendix A.0.2) has also been used to implemented to estimate (23).

Net hazard λ_{Cancer} will be modelled using sex, country or center, birth year and age at diagnosis as covariates. Once again, baseline hazard will be modelled as the exponential of a degree 3 B-spline, with two knots at the 1/3 and 2/3 quantiles of the event times, as well as the effect of birth year and age of diagnosis, where knots will be place at quantiles 1/3 and 2/3 of the corresponding distributions. The same variable selection procedure will be implemented, and the expression of net hazard can be written as:

$$\lambda_{\text{Cancer}}(t;x) = \underbrace{\boxed{\lambda^0_{\text{Cancer}}(t)}}_{\text{B-spline}} \exp \sum_{i=1}^{p=4} \left(\alpha_i(t) g_i(x_i) + \frac{1}{2} \sum_{j=1, j\neq i}^{p=3} \alpha_{j,i}(t) g_j(x_j) g_i(x_i) \right), \tag{24}$$

where t is the time since cancer diagnosis, x_1 the sex, x_2 the country or center, x_3 the birth year and x_4 the age at diagnosis. Sex and country are modelled with a linear effect, and thus $g_1 = g_2 = \mathrm{I}$, and g_3 and g_4 correspond to the B-splines used to model birth year and age at diagnosis respectively.

Additional covariates may be added to the model. In this case, the procedure will be analogous: continuous covariates will be included with a non-linear effect given by a 3 degree B-spline with two knots, categorical covariates will be included with a linear effect, and all the significant interactions and time-dependent effects will be considered.

When performing variable selection, using `glm` software on the equivalent Poisson regression model is recommended, specially when fitting a model to the whole EPIC base. The final model can be then fit with `mexhaz`, avoiding a very time consuming computation. However, building Poisson models requires a tedious previous data management, so a good balance between both methods should be found. In this project, Poisson model was used to estimate λ_{Other} and all the hazard models having age as time scale. Net hazard estimation on cancer cohorts was performed directly with `mexhaz`, and an adapted variable selection procedure was developped to implement on these models. Further details on this issue can be found in Appendix A.0.2.

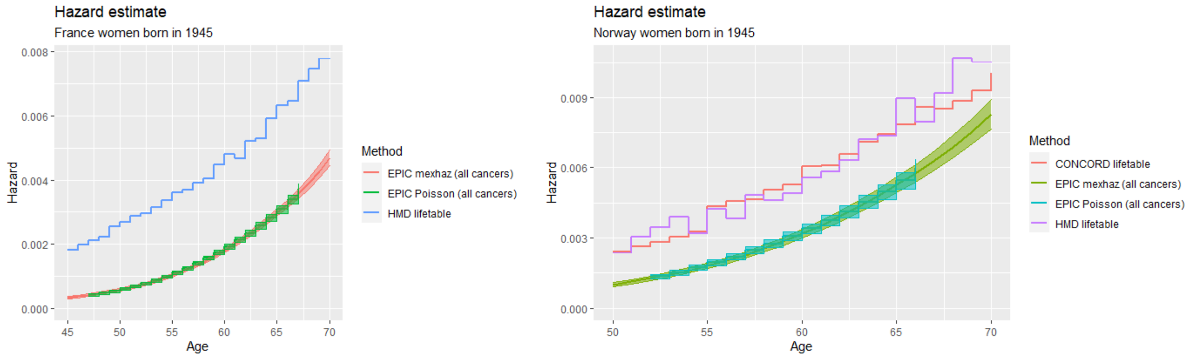
Non-parametric net survival estimation with EPIC based expected hazard has also been implemented using Pohar-Perme estimator. This requires the construction of life tables reflecting EPIC's other causes mortality, for each country or center and sex, which must be passed as an argument to the `R` Pohar-Perme estimator implementation. Life table construction requires previous data management, but it is the only way to obtain non-parametric estimates of relative net survival avoiding selection effects. A fully detailed description of the non-parametric relative survival estimation in EPIC cohort can be found in Appendix B.

Before discussing the obtained results after net survival estimation using the EPIC adapted expected hazard, we will try to assess and quantify the diffence between the EPIC and the general population all-causes mortality, which will give a more precise idea of the magnitude of the selection effect we were trying to avoid. To do so, we will estimate EPIC's overall hazard using (23) including also full follow-up times of individuals diagnosed with the cancer of interest, allowing therefore an accurate comparison of both populations' mortality. Further details and the corresponding results are showed in the next section.

3.1.3 EPIC and general population expected hazard comparison

In order to better illustrate the non-comparability between general and EPIC population, all-causes (overall) hazard will be computed for all individuals in EPIC cohort and for the population obtained from HMD and/or CONCORD life tables. As we mentionned above, EPIC hazard will be estimated using both the equivalent `mexhaz` and Poisson regression models, given by (23), and the general population hazard will be obtained directly from life tables, where mortality rates are provided.

Figure 9 illustrates the results for French (a) and Norwegian (b) women born in 1945, corresponding to the colorectal cancer registries where selection effect was respectively more and less drastic. Overall hazard was computed for HMD and EPIC population, using `mexhaz` and the equivalent Poisson regression model for the latter. As a CONCORD life table is also available for the Norwegian population, the corresponding estimate is also included in that case.



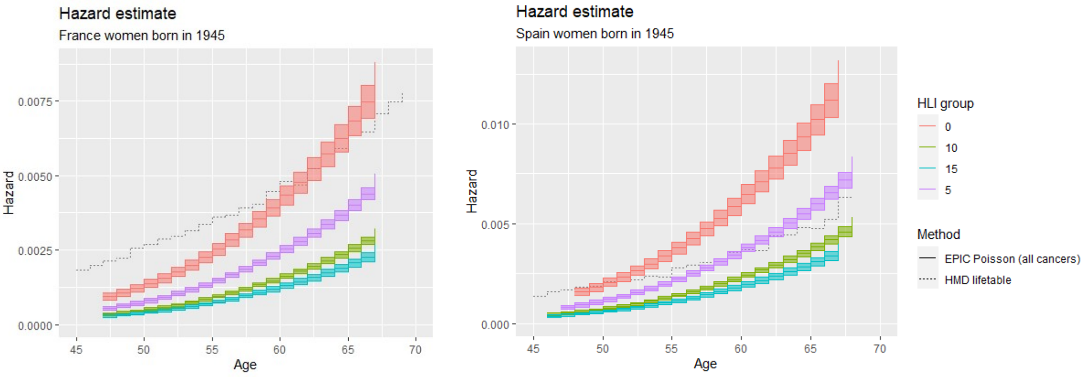
(a)

(b)

Figure 9 – EPIC's and general population's overall hazard for French (a) and Norwegian (b) women born in 1945.

The results explicitly account for the already mentionned selection effect. General population mortality overestimates in both cases EPIC's overall mortality, and thus both populations are not comparable in terms of other causes mortality, as the proportion of individuals diagnosed with cancer is negligible. On the other hand, the magnitude of the difference between both curves reflects the amplitude of the selection effect. As we mentionned above, French women showed the more dramatic inconsistencies while estimating Pohar-Perme's relative net survival using general population life tables, and so the difference between hazards on Figure 9a is considerably high. Figure 9b accounts for the reciproc effect. A more detailed analysis could be envisaged, aiming for instance to define a selection effect indicator based on the difference between the curves in Figure 9. This could be done by computing the supremum norm of the difference at each age value, or by measuring the area between the curves.

What we will try to do here is to have an idea of why general population hazard overestimates EPIC's one. One first hypothesis is that this may be caused by the *healthy bias* of EPIC cohort, which means that EPIC individuals have a healthier lifestyle than general population. In order to assess this, we need to re-estimate overall hazard including covariates which account for individuals' lifestyle. We will use the Healthy Lifestyle Index (HLI), a score defined by McKenzie et al. [27] in order to investigate the joint effect of modifiable factors on the risk of cancer. It is defined as a composite measure reflecting information on diet, physical activity, smoking, alcohol consumption and anthropometry. HLI is available on EPIC cohort as both a continuous and a categorical variable, taking values from the worst lifestyle score, 0, to the better, 20, for the former case, and within four ordered categories for the latter. Two models will be computed, adding continuous or discrete HLI to (23) and the significant time-dependent effects and interactions. HLI models were implemented as Poisson regression models, to avoid time consuming computation. The resulting estimates will be compared to general population hazard.



(a) (b)

Figure 10 – EPIC's and HMD's overall hazard for French (a) and Spanish (b) women born in 1945. EPIC's hazard estimates are depicted for the four Healthy Lifestyle Index (HLI) categories, and have been modelled by a Poisson regression model.

As continuous and discrete versions of HLI showed similar results, the estimates of the model using categorical HLI are illustrated for simplicity. Figure 10 displays hazard estimates for French (a) and Spanish (b) women born in 1945 of EPIC and HMD populations. EPIC estimates are depicted for the four HLI categories, where 0 represents the worst lifestyle and 15 the best one. Figure 10b shows a common trend among all countries, sexes and birth years, and Figure 10a represents a special case. Starting with the former, general population hazard is situated as expected between the two intermediate lifestyle categories. EPIC individuals with the highest HLI show a lower overall hazard than general population, and their curve follows the same pattern as the HLI-free estimate; always underestimating general population mortality. For the French cohort (Figure 10a), general population is closer to EPIC individuals with the lowest HLI score. French women cohort, which showed the more drastic selection effect when computing relative survival with general population, appears as significantly healthier than HMD population. However, we are sceptical about assuming that French general population lifestyle corresponds to the worst HLI score, as Figure 10a might be showing. The mentionned *healthy bias* may not explain -even partially- the difference observed in Figure 9a, and a further analysis to understand French situation should be carried out.

3.2 Cause-specific survival estimation

Cause-specific estimation will be mainly performed using Charvat et al.'s regression model [15]. As we did in the relative survival framework, cause-specific non-parametric estimation has also been considered (see Appendix B), but we will here focus on parametric estimation via `mexhaz`. Net hazard will be directly estimated using the available underlying causes of death (see Section 1.2.1), and it will be firstly modelled using birth year, sex, country or center and age at diagnosis as covariates in the same way as we did to model relative hazard in (24).

3.3 Results

We present in this section the results of net survival estimation in the EPIC cohort using both cause-specific and relative survival approaches. We focus on cancer types with high mortality rates, and whose survival is influenced by external risk factors, as one of our main objectives was to assess if adjustment by those additional covariates would have a significant influence on net survival estimation. This would be particulary interesting in the relative survival framework, as we would be able to compare mortality among cancer and non-cancer patients accounting for the effect of determinant exposures as tobacco smoking. This is the case of lung cancer, which is the most common cancer type in terms of incidence and mortality, and who is mainly caused by tobacco smoking [1]. The proportion of smokers among patients diagnosed with lung cancer is higher than among general population, which makes the comparison of their other causes mortality inaccurate if tobacco information is not taken into account, as smoking is also a risk factor for other causes of death. As information on tobacco smoking is available for EPIC individuals, we will be able to adjust survival estimates using smoking related covariates, and to asses its influence on cause-specific and relative survival estimation. We will also estimate net survival for colorectal cancer patients, whose survival is higher and it is also influenced by lifestyle factors. Alcohol consumption, tobacco smoking and consumption of processed meat have been shown as colon and rectum carcinogenic agents [1], so the influence of the already mentionned Healthy Lifestyle Index [27] on net survival estimation will be assessed.

We first present cause-specific and relative survival lung cancer estimates, computed with the flexible regression model `mexhaz` according to (24). Net survival was first modelled using sex, country, age at diagnosis and birth year as covariates, and then adding information on tobacco smoking through a categorical covariate distinguishing between never, former and current smokers. Two main trends were found among all countries, sexes, birth years and ages of diagnosis, and they are illustrated by two representative examples on Figure 11. Relative and cause-specific survival estimates are depicted when the smoking status is not taken into account (left column) and when it is included in the model (right column).

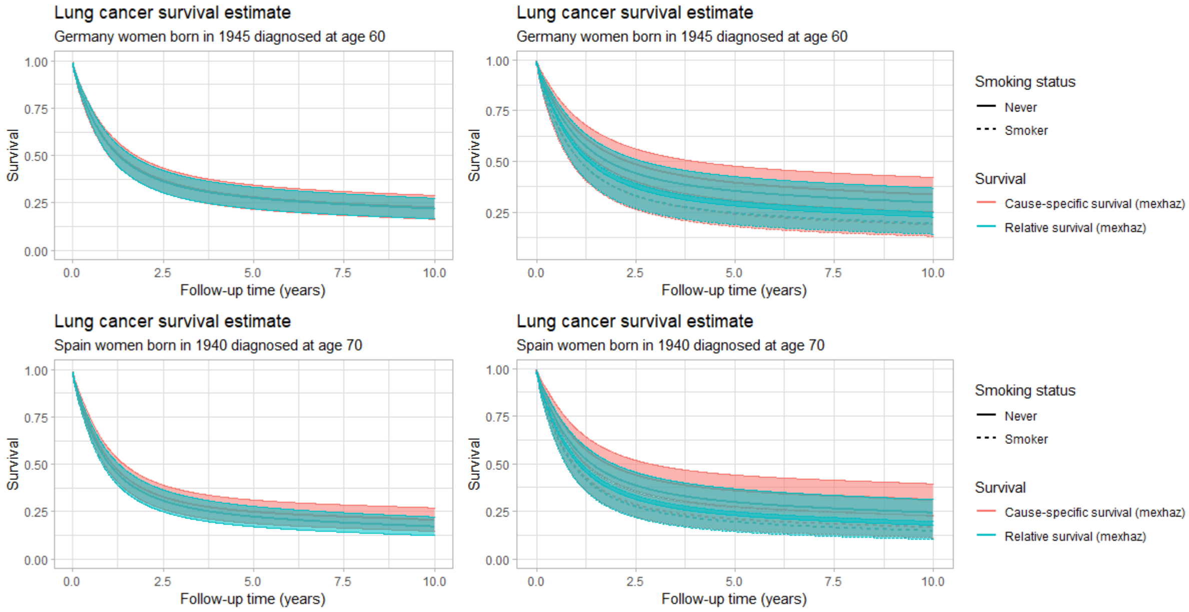


Figure 11 – Relative and cause-specific survival estimates for German (first row) and Spanish (second row) women diagnosed with lung cancer. Smoking status has not been taken into account for the curves depicted in the left column, and it has been included in the model for the second column estimates.

First row on Figure 11 represents the first trend found among all EPIC subpopulations, exemplified here in the case of German women born in 1945 and diagnosed with lung cancer at 60 years old. When smoking is not taken into account, both relative and cause-specific approaches provide overlapping curves. When smoking information is considered, only the estimates for current smokers stay overlapped, and a small separation (within the confidence intervals) for never smokers appear. The second row shows the case where cause-specific and relative survival curves are not overlapped when the smoking covariate is not considered, but cause-specific survival takes higher values for all follow-up times. Even if this discrepancy appears, both curves coincide within the confidence intervals. This trend is represented by Spanish women born in 1940 and diagnosed with lung cancer at the age of 70. When the smoking status is included in the model, cause-specific and relative survival curves do overlap for current smokers, but the difference remains -and even expands- for never smokers. When differences appear, both curves coincide within the confidence intervals, and cause-specific approach takes higher values in all EPIC subpopulations.

Colorectal cancer estimates are presented analogously. Net survival was first modelled according to (24) using country, sex, birth year and age at diagnosis as covariates. Healthy Lifestyle Index (HLI) was then included in the model with a non-linear effect given by a degree 3 B-spline with two knots at the 1/3 and 2/3 corresponding quantiles. We present in Figure 12 two representative examples of net survival estimates. The first column illustrates the curves produced by the models that did not include HLI as a covariate, and the second the ones produced by the models that did.

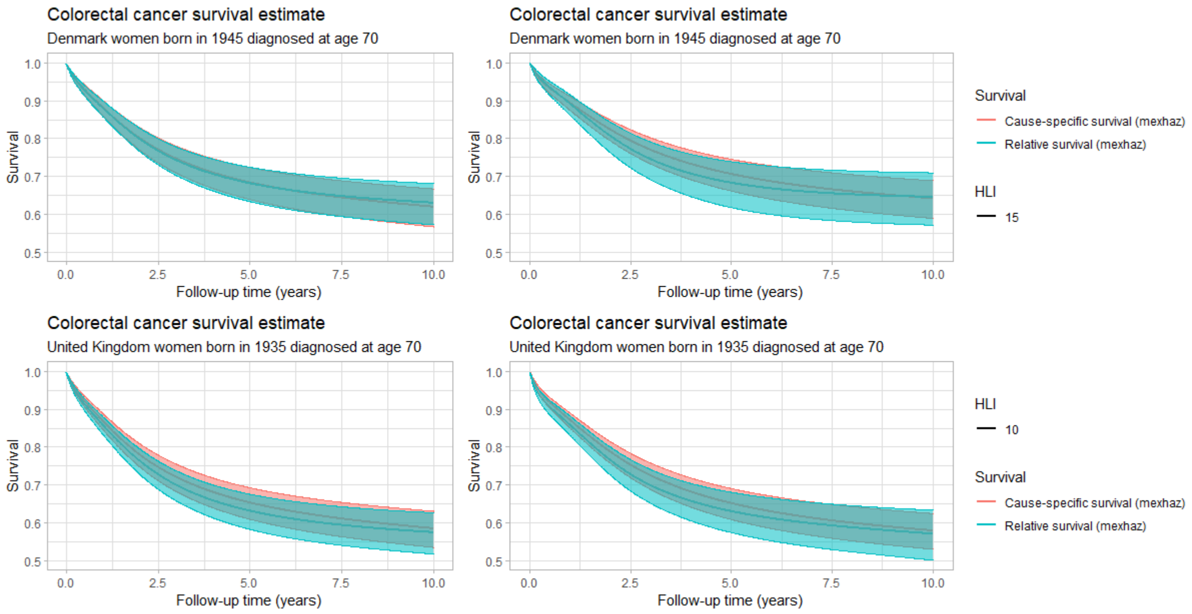


Figure 12 – Relative and cause-specific survival estimates for Danish (first row) and British (second row) women diagnosed with lung cancer. Healthy Lifestyle Index has not been taken into account for the curves depicted in the left column, and it has been included in the model for the second column estimates.

When Healthy Lifestyle Index is not included in the model, trends among all EPIC subpopulations can be represented by the two examples shown in Figure 12. Danish women born in 1945 and diagnosed with colorectal cancer at the age of 60 represent the case where cause-specific and relative survival curves overlap. Small discrepancies are sometimes found in cases such as the British cohort of women born in 1935 and diagnosed at 70 years old with colorectal cancer. Cause-specific curves always take higher values than relative survival estimates, for all the follow-up times, and both curves always coincide within the confidence intervals. The main difference between colorectal and lung cancer trends in results appears when adjusting by the additional covariate. In this case, adding the Healthy Lifestyle Index does not decrease, but enlarge the existing discrepancy between the two approaches, whose magnitude depend on the subpopulation and the considered HLI value.

3.4 Discussion

Cause-specific and relative survival approaches have been implemented in the EPIC cohort in order to estimate lung and colorectal cancers net survival. Several methods have been assesed, and parametric estimation via the flexible regression model `mexhaz` has been considered the most appropriate for our study. Overall results showed overlapping, or coincident within the confidence intervals, cause-specific and relative survival curves. When differences appeared, cause-specific estimates took higher values for all follow-up times.

Adjusting for additional covariates had a different effect on lung and colorectal cancer. In the case of the former, adding smoking status to the model did correct discrepancies between the two approaches for smoker patients, but had no effect, or enlarged differences, for never and former smokers. Adjusting by Healthy Lifestyle Index when estimating colorectal cancer net survival did not decrease, but sometimes broadened discrepancies between cause-specific and relative survival curves. However, all discrepancies we are here considering remain within the confidence intervals of both curves, so we must be cautious and do not state that adjusting by additional covariates produced a significant reduction of differences when they appeared.

After analysing the results, we can state that both cause-specific and relative survival methods provide net survival estimates not significantly different. This suggest that causes of death information, whose reliability is questionable for some authors, can provide comparable estimates of net survival to those produce in the relative survival framework. We may then recommend the use of cause-specific approach due to the simplicity of its implementation, in contrast to the relative survival one, which needs a comparable population. Relative survival should also be computed when this population is available, as both methods providing overlapping or non-significantly different survival curves will suggest the high quality of the estimates.

It is important to keep in mind that EPIC is a highly-selected population, so the obtained net survival estimates might not be generalizable to the general population. This has been illustrated through the comparison of estimates of hazards for mortality derived in EPIC vs those present in the HMD/CONCORD life tables, which allowed the quantification of the selection effect. Nonetheless, the produced estimates will be useful for internal comparison within the EPIC cohort (e.g. to compare net survival of cancer patients with or without history of comorbidity).

We would like to underline that this project has provided cancer net survival estimates at European level, implemented by both cause-specific and relative survival approaches. All the details on methodology, the developped `R` tools, the produced models and life tables, and the results (via a `R shiny` application) will be available for IARC internal usage, hoping they will be useful for the projects interested on net survival estimation.

A Further estimation methods of net survival

A.0.1 Non-parametric estimation

Non-parametric methods of estimation keep their interest as they provide a first look at the data without assumptions about the effect of covariates needed, and relate the group experience instead that the invidual level. Same measures are interesting in both relative and cause-specific setttings, but methods of estimation are different, since different information is available in the data.

The non-parametric estimation of the overall survival function, an estimate of the net survival function in the cause-specific setting, can be carried out using the Kaplan-Meier standard method. As it was mentionned, special attention must be paid when the other death censoring is informative, which leads to biased survival estimates. This may happen, for instance, when both the population and the cancer hazard are afected by a common set of covariates. A solution to this problem was proposed by Satten et al. [28] and Robins [29], consisting on a weighted version of the Nelson-Aalen estimator of cumulative hazard. However, the introduced weighting requires the survival probabilities of a comparable external population or life table. Thus, using this method would suppose having fo face both the weak points of cause-specific and relative survival, as a comparable external population has to be found and the corresponding mortality estimates computed. This calculation seems not only unwanted but unneeded, as pertinent parametric methods of cause-specific survival which require just the cancer population are available, and will be presented in the next section.

The first estimators which have widely been used in the relative survival framework are the Ederer I [30], Hakulinen [31] and Ederer II [30] estimators. However, Pohar Perme et al. [11] recently showed that these approaches either are not well defined estimators of net survival or produce biases and/or inconsistencies. They also proposed a new estimator that does not require modeling and which consistently estimates relative net survival. To introduce it from a more intuitive point of view, let's present it as a correction of Ederer II estimator, which estimates net hazard as:

$$d\hat{\Lambda}^{EII}_{Cancer}(t) = \frac{N(t)}{Y(t)} - \frac{\sum_{i=1}^{n} Y_i(t) d\Lambda_{P_i}(t)}{Y(t)}. \tag{25}$$

The hazard is denoted $d\Lambda$ as it corresponds to a picewise constant hazard constant on the intervals which limits are defined by the end of follow-up times of the individuals. The first term of the difference estimates the overall hazard as the number $N(t)$ of events at time t over the number $Y(t)$ of individuals at risk at time t. The second term is an estimate of the expected hazard λ_{Other} at time t only if the used life table or external population is comparable to the cancer cohort. The quantity $d\Lambda_{P_i}(t)$ is the probability of dying from other causes at time t of an invidual of the external population with the same demographic characteristics (sex, country and birth year when using a life table) as the i-th individual of the cohort. If that individual is at risk at time t (i.e. $Y_i(t) = 1$), he or she will contribute to the λ_{Other} hazard according to the probability of dying from other causes at time t of an external individual with his/her same demographic conditions. However, even if the external population is comparable to the cancer cohort, Ederer II produces a biased estimate of net survival [11]. The newly proposed estimator corrects this problem by weighting the counting process according to the general population survival distribution:

$$d\hat{\Lambda}^{PP}_{Cancer}(t) = \frac{\sum_{i=1}^{n} N_i^w(t)}{\sum_{i=1}^{n} Y_i^w(t)} - \frac{\sum_{i=1}^{n} Y_i^w(t) d\Lambda_{P_i}(t)}{Y^w(t)}, \tag{26}$$

where

$$N_i^w(t) = \frac{N_i(t)}{S_{P_i}(t)}, \qquad Y_i^w(t) = \frac{Y_i(t)}{S_{P_i}(t)}, \tag{27}$$

$N_i(t)$ equals 1 if the i-th individual dies in the interval containing t and 0 otherwise, and $S_{P_i}(t) = \mathrm{P}_{P_i}(T > t)$ is the probability of dying from other causes after time t of an individual of the external population with the same (demographic) characteristics as the i-th individual of the cohort. This weighting reproduces the exposure time observed in the hypothetical world where cancer would be the only cause of death. Because of the other cause mortality, the number of at risk individuals in the real world is lower than at the hypothetical one. The Pohar-Perme estimator uses the probabilites S_{P_i} to reinforce the effect of deaths and risk indicators of those individuals with a high probability of dying due to other causes in the real world, as we would have much more individuals with their same characteristics in the hypothetical one.

If the external population is comparable to the cancer cohort, (26) is an unbiased and consistent estimator of net survival, which we will use to estimate non-parameric relative survival. An `R` implementation of Pohar-Perme estimator is available in the `relsurv` package [32].

A.0.2 The equivalent Poisson regression model

The main disadvantage of the `mexhaz` model lies on the computation times of its `R` implementation, which enormously increase with the complexity of the model and the size of the dataset. Even so, estimation with `mexhaz` is strongly recommended in most cases due to the remarkable simplicity of its implementation and the accuracy of the results. However, one may prefer in some situations to sacrifice the mentionned simplicity and gain in computation time, specially when implementing very complex models on large datasets. Holford [35] and Laird and Olivier [36] first prooved that the likelihood of a piecewise exponential hazard model was equivalent to the one of a certain Poisson regression model. As we will briefly illustrate, this can also be generalised to a more flexible hazard model as the `mexhaz` one. This approach is based on the generation of pseudo-observations by splitting on intervals the follow-up time. This may considerably increase the size of the dataset, but provides some other computational advantages, as fitting hazard models using the standard `glm` software [3] and having access to its variable selection procedures (which, for instance, do not still have a `mexhaz` counterpart).

Let's first consider fitting a proportional hazards model of the usual form

$$\lambda_i(t; x_i) = \lambda_0(t) \exp(x_i\beta), \tag{28}$$

for the i-th individual of the cohort with covariates x_i. We then partition time axis into J intervals with cutpoints $0 = \tau_0 < \tau_1 < \cdots < \tau_J = \infty$, being $[\tau_{j-1}, \tau_j)$ the j-th interval. We will assume that the baseline hazard is constant within each interval: $\lambda_0(t) = \lambda_j \; \forall \, t \in [\tau_{j-1}, \tau_j)$, and then model the baseline hazard with J parameters, representing the risk for the reference individual in one particular interval. We may now rewrite (28) as

$$\lambda_{ij} = \lambda_j \exp(x_i\beta), \tag{29}$$

where λ_{ij} is therefore the hazard corresponding the i-th individual in j-th interval. λ_j is the baseline hazard for interval j and $\exp(x_i\beta)$ the relative risk for an individual with covariates x_i compared to the baseline at any time. The expression (29) is equivalent to the model

$$\log \lambda_{ij} = \log \lambda_j + x_i\beta, \tag{30}$$

which is as standard log-linear model where time categories are treated as a factor. Let now t_{ij} be the time lived by the i-th individual in the j-th interval and δ_{ij} the indicator of i-th invidual dying in interval j. Then, the piecewise exponential model (30) can be fitted to the data by treating the death indicators d_{ij} as they were independent Poisson observations with means

$$\mu_{ij} = t_{ij}\lambda_{ij}. \tag{31}$$

Taking logs on (31) we obtain

$$\log \mu_{ij} = \log t_{ij} + \log \lambda_j + x_i\beta, \tag{32}$$

and thus, the piecewise exponential proportional hazards model is equivalent to a Poisson log-linear model for the pseudo-observations, one for each combination of individual and interval, where the death indicator is the response and the log of exposure time enters as an offset. For a proof of the likelihood equivalence one may see [37] or [34]. These model can be extended to introduce interactions and time-dependent effects, and be fitted using `glm` functions.

One can also group observations according to the covariates values and add up the measures of expousure and the death indicators. We may define d_{ij} as the number of deaths and t_{ij} as the total exposure of individuals with characteristics x_i in interval j. The estimates, standard errors and likelihood ratio tests would be exactly the same as for individual data [37]. This is the approach we will follow in our case.

The Poisson regression model can be computed to obtain an equivalent estimate of `mexhaz` hazard if time divisions are smooth enough. However, as we mentionned before, its implementation requires further data management and we will make use of it only in specific situations as lifetable building, where this method might be useful.

B Non-parametric net survival estimation in the EPIC cohort

Non-parametric estimation of net survival provides a first look at the data without making any assumption about the effect of covariates needed. We present in this Appendix the implemented methodology and an overview of the results after computing both cause-specific and relative non-parametric survival in the EPIC cohort.

Cause-specific survival was computed using the Kaplan-Meier estimator. Deaths due to cancer were considered the end-point of interest, and deaths due to other causes were censored (see Section 1.2.1). The Pohar-Perme estimator introduced in Appendix A.0.1 was used to estimate non-parametric relative survival. The main problem we faced at this stage was the one of building a life table adapted to the EPIC cohort, as the general population ones can not be used to estimate λ_{Other} (see Section 3.1.1). We envisaged two approaches in order to build an adapted life table. The first relied on the idea of keeping our net survival estimation completely non-parametric, and thus to estimate λ_{Other} without asuming any model. The second admitted a semi-parametric relative survival estimation, as λ_{Other} was computed via a Poisson regression model (see Appendix A.0.2).

The non-parametric life table was conceived in the setting of Lexis diagrams, which allow the representation of survival times in a two dimensional time space. Figure 13 displays an example of a Lexis diagram, where the abscissa represents the calendar year and the ordinate the age of individuals. For an individual recruted at age a_r and at calendar year y_r, and leaving the study at age a_e and at calendar year y_e, his or her life line is represented by the line segment having (y_r, a_r) and (y_e, a_e) as end-points. Both age and calendar year are considered continuous and take values in $\mathbf{R}^+$.

In order to build a life table we need to estimate the mortality rates for the given population, i.e. the probability of dying at age a and at calendar year y for all ages $a \in \mathbf{Z}^+$ and all calendar years $y \in \mathbf{Z}^+$. If we denote by T_A the age at death and by T_Y the calendar year at death, we have to estimate the quantity:

$$q_{ay} = \mathrm{P}(a < T_A < a+1\,,\, y < T_Y < y+1 \,|\, T_A \geq a\,,\, T_Y \geq y). \tag{33}$$

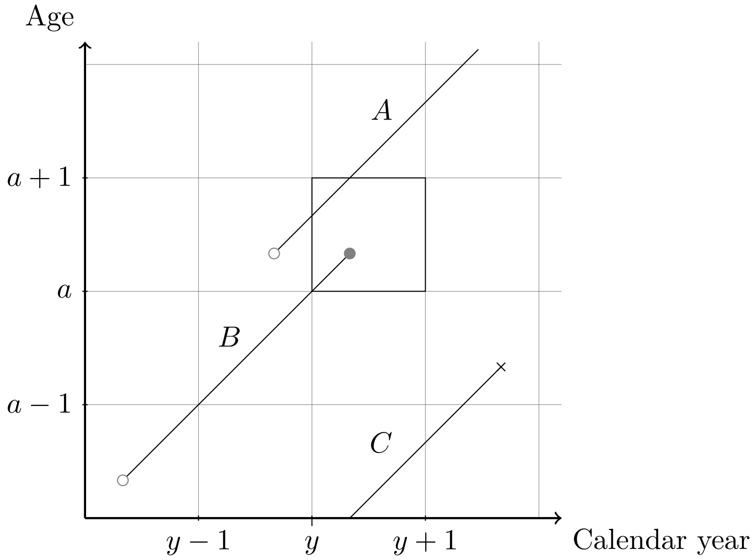


Figure 13 – Lexis diagram illustrating survival times with right censoring (cross) and left truncation (empty circle). Events are represented by filled circles.

In the Human Mortality Database Methods Protocol, Wilmoth et. al [38] suggested that when follow-up times are available, the best way to estimate (33) is through the ratio:

$$\hat{q}_{ay} = \frac{\text{Number of events in } [y, y+1] \times [a, a+1]}{\text{Exposure time in } [y, y+1] \times [a, a+1]}. \tag{34}$$

This ratio is not, however, a well-defined estimator of the mortality rates (33), as it can take values higher than 1 when, for instance, the given square contains only events of interest. We have anyway implemented (34) to build the life tables, as we have enough individuals in the risk set to always avoid this problem. Anyhow, a well-defined non-parametric estimator should be found in order to properly estimate mortality rates. Some new estimators were proposed, but no conclusive results were found ensuring their good properties.

The non-parametric estimator (34) was implemented for all non-empty Lexis squares (i.e. the squares $[y, y+1] \times [a, a+1]$ of Figure 13 with non-empty intersection with the life lines) of all EPIC subpopulations given by all country-sex combinations. The computed estimates were organized in a life table, providing the mortality rates for all ages and calendar years in the given country and sex.

A Poisson regression model was also implemented to obtain a 'parametric' life table. The number of events and the exposure time in each Lexis square was computed for each combination of country, sex and birth year in the whole EPIC population, and then the regression model was fitted, modelling the effect of birth year as a degree 3 B-spline with two knots at the 1/3 and 2/3 quantiles (see Appendix A.0.2 for details). If we assume that λ_{Other} is a piecewise constant function along the age scale, the resulting hazard estimates λ_{ay} for each age a and each calendar year y can be transformed into mortality rates through the expression:

$$q_{ay} = 1 - \exp(-\lambda_{ay}), \tag{35}$$

which can be easily derived under the mentionned assumption.

We present now an overview of the results of cause-specific and relative survival non-parametric estimation for the EPIC patients diagnosed with colorectal cancer. Survival curves were computed for each combination of country and sex, and stratified by age at diagnosis groups given by the quantiles 0.2, 0.4, 0.6 and 0.8. Figure 14 illustrates two representative examples of the survival curves for the whole group (left column) and of the estimates for each strata at 5 years of follow-up (right column).

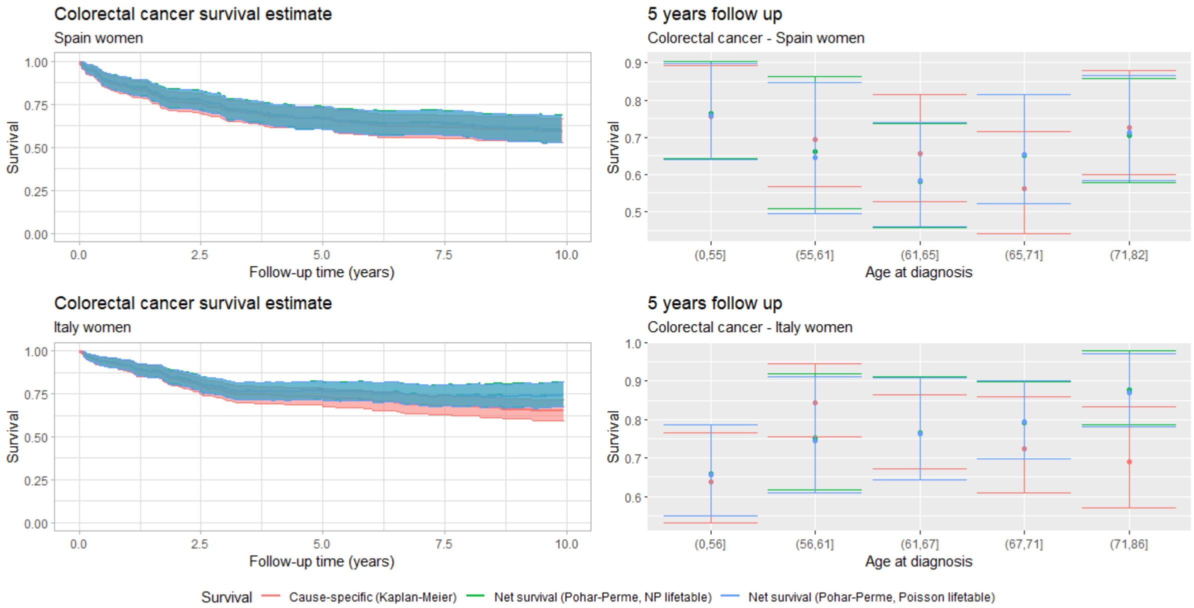


Figure 14 – Non-parametric cause-specific and relative survival estimates for the Spanish women (top) and Italian women (bottom) cohort diagnosed with colorectal cancer. The left column depicts the survival curves for the entire group and the right column shows the net survival estimates at 5 years of follow-up for each age at diagnosis strata.

First row on Figure 14 displays the survival curves and estimates for Spanish women diagnosed with colorectal cancer. Cause-specific and relative survival curves overlap within the confidence intervals, as well as the estimates at 5 years of follow-up. Second row depicts the survival curves and estimates for Italian women diagnosed with colorectal cancer. In this case, relative survival takes higher values than cause-specific after 2.5 years of follow-up, when it stabilizes and stop decreasing. This trend might be explained if the estimated λ_{Other} would be overestimating the real mortality due to other causes of this subpopulation.

From Figure 14, one may underline the fact that the completely non-parametric and the semi-parametric relative survival estimates do overlap within the confidence intervals, for all countries and sexes and when stratifying by age at diagnosis groups. This suggest that a completely non-parametric estimation of net survival in the relative survival framework is viable, and may be equivalent to a semi-parametric approach, where information on sex, country and birth year of the whole EPIC population is used to estimate λ_{Other}. On the other hand, we have to recall that non-parametric estimates only account for the group experience, and provide a first look at the data before implementing more precise models, which can produce estimates at the individual level. No conclusive results should be taken from this first overview, and the analysis of the parametric estimation via the flexible regression model `mexhaz` is recommended before stating any solid conclusion (see Section 3.3).

References

[1] Wild C.P., Weiderpass E., Stewart B.W., editors (2020). *World Cancer Report: Cancer Research for Cancer Prevention.* Lyon, France: International Agency for Research on Cancer. Available from: http://publications.iarc.fr/586. Licence: CC BY-NC-ND 3.0 IGO.

[2] Dickman, P.W., Auvinen, A., Voutilainen, E.T., Hakulinen, T. Measuring social class differences in cancer patient survival: it is neccessary to control for social class differences in general population mortality? A Finnish population-based study. *J Epidemiol Community Health* 1998; **52**, 727-34.

[3] Dickman, P.W., Sloggett, A., Hills, M., Hakulinen, T. Regression models for relative survival. *Stat Med* 2004; **53**, 51-64.

[4] Dickman, P.W., Adami, H.O. Interpreting trends in cancer patient survival. *J Intern Med* 2006; **206**, 103-17.

[5] Rachet, B., Woods, L.M., Mitry, E. *et al.* Cancer survival in England and Wales at the end of the 20th century. *Br J Cancer* 2008; **99**, S2-S10.

[6] Allignol, A., Schumacher, M., Beyersmann, J. *Competing Risks and Multistate Models with R.* Springer, New York, 2010.

[7] Andersen, P., Borgan, ∅., Gill, R., Keiding, N. *Statistical Models Based on Counting Processes.* Springer, New York, 1993.

[8] Aalen, O., Borgan, ∅., Gjessing, H. *Event History Analysis.* Springer, New York, 2008

[9] Kalbfleisch, J.D., Prentice, R.L. *The Statistical Analysis of Failure Time Data* John Wiley, New York, 1980.

[10] Kaplan, E.L., Meier, P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*, 1958; **53**, 457-81.

[11] Pohar Perme, M., Stare, J., and Estève, J. On Estimation in Relative Survival. *Biometrics*, 2012; **68**, 113-120.

[12] Sarfati, D., Blakely, T., Pearce, N. Measuring cancer survival in populations: relative survival vs cancer-specific survival. *International Journal of Epidemiology* 2010.

[13] World Health Organization. *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death, based of the recommendations of the ninth revision conference, 1975.* Geneva: WHO, 1977.

[14] Ezzati, M., Lopez, A.D. Estimates of global mortality attributable to smoking in 2000. *Lancet* 2003; **362**, 874-52.

[15] Charvat, H., Remontet, L., Bossard, N., Roche, L., Dejardin, O., Rachet, B., Launoy, G., Belot, A. and the CENSUR Working Survival Group. 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* 2016; **35**, 3066-3084.

[16] Esteve, J., Benhamou, E., Croasdale, M., Raymond, L. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine* 1990; **9**(5), 526-538.

[17] Bolard, P., Quantin, C., Abrahamowicz, M. Esteve, J., Giorgim R., Chadha-Boreham, H., Binquet, C., Faivre, J. Assesing time-by-covariate interactions in relative survival models using restricted cubic spline functions. *Journal of Cancer Epidemiology and Prevention* 2002; **7**(3), 113-122.

[18] Giorgi R., Abrahamowicz M., Quantin C., Bolard P., Esteve J., Gouvernet J., Faivre J. A relative survival regression model using B-spline functions to model non-proportional hazards. *Statistics in Medicine* 2003; **22**(17), 2767–6784.

[19] Remontet L., Bossard N., Belot A., Esteve J. An overall strategy based on regression models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine* 2007; **26**(10), 2214–2228.

[20] Royston P., Ambler G., Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 1999; **28**(5), 964–974.

[21] Nelson C.P., Lambert P.C., Squire I.B., Jones D.R. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine* 2007; **26**(30), 5486–5498.

[22] Charvat, H. and Belot, H. mexhaz: Mixed Effect Excess Hazard Models. R package, version 1.7. 2019. https://CRAN.R-project.org/package=mexhaz

[23] Dennis JE, Schnabel RB. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall: Englewood Cliffs, 1983.

[24] International Agency for Research on Cancer. EPIC study. 2020. Retrieved from https://epic.iarc.fr/.

[25] Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on 23 March 2020).

[26] Di Carlo V, Rachet B, Bannon F, Woods LM, Maringe C, Bonaventure A, Coleman MP, Allemani C. Life tables for the CONCORD programme. Available from: http://csg.lshtm.ac.uk/life-tables (downloaded on 11 June 2020).

[27] McKenzie, F., C. Biessy, P. Ferrari, H. Freisling, S. Rinaldi, V. Chajès, C. C. Dahm, et al. 2016. "Healthy Lifestyle and Risk of Cancer in the European Prospective Investigation Into Cancer and Nutrition Cohort Study." Medicine 95 (16): e2850. doi:10.1097/MD.0000000000002850. http://dx.doi.org/10.1097/MD.0000000000002850.

[28] Robins, J.M. 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*, pp. 24-33. Alexandria, Virginia, U.S., 1993.

[29] Satten, G.A., Datta, S., and Robins, J. Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters* 2001; **54**, 397-403.

[30] Ederer, F., Axtell, L.M., and Cutler, S. J. *The Relative Survival Rate: A Statistical Methodology*, pp. 101-121. Bethesda, Maryland, U.S. National Cancer Institute Monograph 6, 1961.

[31] Hakulinen, T. Cancer survival corrected for heterogeneity in patient withdrawal. *Biometrics*, 1982; **38**, 933-942.

[32] Pohar Perme, M., Pavlic, K. Nonparametric Relative Survival Analysis with the R package relsurv. *Journal of Statistical Software*, 2018; **87**(8), 1-27.

[33] Fox, J. *Applied Regression Analysis and Generalized Linear Models*. Sage, Thousand Oaks, CA, 3rd edition, 2016.

[34] Fauvernier, M. Splines multidimensionnelles pénalisées pour modéliser le taux de survenue d'un événement. Application au taux de mortalité en excès et à la survie nette en épidémiologie des maladies chroniques. 2019, PhD Thesis. Université Claude Bernard Lyon 1, Lyon.

[35] Holford, T.R. The Analysis of Rates and of Survivorship Using Log-Linear Models. *Biometrics* 1980; **36**, 209-305.

[36] Laird, N., Olivier, D. Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. *Journal of the American Statistical Association* 1981; **76:374**, 231-240.

[37] Rodríguez, G. Lecture Notes on Generalized Linear Models. 2007. Retrieved from http://data.princeton.edu/wws509/notes/.

[38] Wilmoth, J.R., Andreev, K., Jdanov, D., Glei, D.A. and Riffe T. with the assistance of Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., Vachon, P., Winant, C., Barbieri, M. Methods Protocol for the Human Mortality Database. Last Revised: October 5, 2019 (Version 6).