Competing risks models with two time scales

— Extended abstract —

Angela Carollo
carollo@demogr.mpg.de
Max Planck Institute for Demographic Research
Rostock, Germany
Leiden University Medical Center, Leiden, The Netherlands

Hein Putter
h.putter@lumc.nl
Leiden University Medical Center, Leiden, The Netherlands

Paul H.C. Eilers
p.eilers@erasmusmc.nl
Erasmus University Medical Center, Rotterdam, The Netherlands

Jutta Gampe
gampe@demogr.mpg.de
Max Planck Institute for Demographic Research
Rostock, Germany

November 1, 2023

Abstract

Competing risks models can involve more than one time scale. A relevant example is the study of mortality after a cancer diagnosis, where time since diagnosis but also age may jointly determine the cause-specific hazards of death. Here death due to cancer and other causes of death are the competing risks. Another example is transition out of non-marital cohabitation where age of the individual and duration of cohabitation both are relevant time scales. Here the competing risk are transition to marriage or separation.

Multiple time scales have rarely been explored in the context of competing events. Here, we propose a model in which the cause-specific hazards vary smoothly over two times scales. It is estimated by two-dimensional P-splines, exploiting the equivalence between hazard smoothing and Poisson regression. The data are arranged on a grid so that we can make use of generalized linear array models for efficient computations.

As a motivating example we analyse mortality after a breast cancer diagnosis and we distinguish between death due to breast cancer and all other causes of death. The time scales are age and time since diagnosis. We use data from the Surveillance, Epidemiology and End Results (SEER) program. In the SEER data, age at diagnosis is provided with a last open-ended category, leading to coarsly grouped data. We use the two-dimensional penalised composite link model to ungroup the data before applying the competing risks model with two time scales.

Keywords: Cause-specific hazards; Two-dimensional smoothing; P-splines; PCLM; Cancer mortality

1 Introduction

Competing risks describe the situation where individuals are at risk of experiencing one of several types of events [1]. The prototype of a competing risks model is the study of cause-specific mortality.

The building blocks in a competing risks analysis are the cause-specific hazards. They are defined as the instantaneous risk of experiencing an event of a specific type at time t, given no event (of any type) has happened before t. From them the overall survival function, i.e., the probability of no event up to time t, and the cumulative incidence functions, the probability of an event of given type before t, can be derived.

Time is a key quantity in any survival analysis, and it can be recorded over several time scales. For example, after a cancer diagnosis, the risk of death may be studied over time since diagnosis, over age, which is time since birth, or over time since treatment. All time scales progress at the same speed and differ only in their origin. However, each time scale is a proxy for a specific mechanism linked to the event of interest [2]. In the study of mortality after a cancer diagnosis, time since diagnosis measures the cumulative adverse effect of the cancer. Additionally, as individuals age, they become more frail and their capacity of resisting comorbidity deteriorates.

Usually, cause-specific hazards are defined for the same single time scale, and little research has been done on how to handle multiple time scales in competing risks models. Cancer mortality is clearly a function of time since diagnosis, while other causes of death are more naturally modeled over age. Whenever the cause-specific hazards of death for other causes are solely modeled over the time since diagnosis, bias may be introduced in the estimates of the cumulative incidences for cancer mortality [3].

Carollo et al. [4] introduced a model for a single event in which the hazard varies smoothly over two time scales simultaneously and is estimated by tensor products P-splines [5]. Here, we develop this model further for a competing risks setting. Each cause-specific hazard varies over the two time scales, and estimation again is achieved by two-dimensional P-splines smoothing. Therefrom, we calculate the cumulative incidence functions for each cause.

This study is motivated by the analysis of mortality after a breast cancer diagnosis. Breast cancer is the most common cancer diagnosed in women worldwide [6], and the risk of being diagnosed with it increases over age, but varies by race and ethnic group [7], with peaks of diagnosis after age 60 for white women and in the late 40s for women of race/ethnicities other than white.

Mortality rates of women with breast cancer depend on age, race and cancer subtypes [8, 9] however, the relationship between time since diagnosis, age, age at diagnosis and mortality is not yet clear. Although most studies acknowledge the presence of several time scales, they account for them in a suboptimal manner. For example, age at diagnosis is often grouped in broad and arbitrarily-cut categories, and interaction with time since diagnosis is often neglected. Shedding light on the way mortality rates of breast cancer vary with these two time scales, while properly accounting for competing causes is of great empirical and methodological relevance.

We analyse mortality data of women with breast cancer from the Surveillance, Epidemiology and End Results (SEER) registers [10] by time since cancer diagnosis and age. We distinguish between deaths due to breast cancer and all other causes of death. In the SEER data, age at diagnosis is recorded in single years of age up to age 89 and one last open-ended interval, which groups together all diagnosis for individuals older than 89 (age at diagnosis 90+). We employ the two-dimensional penalised composite link model (PCLM) [11] to ungroup the grouped data. Doing so allows us to use the full data provided by SEER, without restrictions on the oldest ages, to estimate the competing risks model.

2 A competing risks model with two time scales

Consider individuals with a diagnosis of cancer. At each point in time, after becoming ill with cancer, an individual is at risk of dying because of the cancer (event of interest) or because of causes other than cancer (competing event). While individuals face the mortality risk from cancer onset, it can be measured only after the cancer is diagnosed. Therefore, in the following, we will focus on the time since diagnosis of the cancer and age at diagnosis rather than time since cancer and age at cancer onset.

Two time scales are relevant for these transitions: The first is the age of the individual, which we indicate with t. The second time scale is time since diagnosis of the cancer, which we denote by s. Because age measures the time since birth, we have that $t > s$ for each individual. Both t and s move at the same speed, meaning that an increment of one unit (for example one day or one month) on the t-scale corresponds to the same increment on the s-scale.

We consider two competing events: Death from beast cancer ($\ell = 1$) and death from other causes ($\ell = 2$). The cause-specific hazards (CSHs) for event type $\ell \in \{1, 2\}$ over the two time scales t and s are defined as

$$\lambda_\ell(t, s) = \lim_{\nu \downarrow 0} \frac{\mathrm{P}\,(\text{ event of type } \ell \in \{t + \nu, s + \nu\} \,|\, \text{no event of any type before } (t, s)\,)}{\nu}. \tag{1}$$

Here $t > s$, so the two-dimensional hazards $\lambda_\ell(t, s)$ are only defined in the lower half-open triangle of $\mathbb{R}^2_+$. They give the instantaneous risk of dying because of cause ℓ for an individual who is alive at age t and s years after receiving the diagnosis of cancer.

The two time scales differ in their origin, which in the case of cancer patients is the age t_0 when the individual is diagnosed with cancer. This age is fixed for each individual but differs among them. In a Lexis diagram with axes t (age) and s (time since diagnosis) individuals move along diagonal lines from $(t_0, 0)$ to $(t_0 + v, v)$ until they leave the risk set (due to event of any kind or censoring). The individual trajectories start at individually different points $(t_0, 0)$.

The same information can also be portrayed in the (u, s)-plane, where u denotes the age at diagnosis and s, as before, is the time since diagnosis. Here individual trajectories run vertically from $(u, s = 0)$ to $(u, s = v)$. We can view the cause-specific hazards equivalently as two-dimensional functions of u and s, $\breve{\lambda}_\ell(u, s)$, where

$$\breve{\lambda}_\ell(u = t - s, s) \equiv \lambda_\ell(t, s), \tag{2}$$

see[4]. The $\breve{\lambda}_\ell(u, s)$ are defined over the full positive quadrant $\mathbb{R}^2_+$.

From the CSHs $\lambda_\ell(t, s)$ or $\breve{\lambda}_\ell(u, s)$, respectively, we obtain the cumulated cause-specific hazards, the overall survival probability and the cumulative incidence functions. In the next section we describe estimation of the CSHs.

2.1 Cause-specific hazards: Model and estimation

We will model (and estimate) the cause-specific hazards $\breve{\lambda}_\ell(u, s)$ over the (u, s)-plane. Back-transformation to the (t, s)-coordinates is straightforward using equation (2). The only assumption that we are going to make about the cause-specific hazards is that they vary smoothly over u and s (and hence over t and s). This will be achieved by two-dimensional P-spline smoothing (tensor products of B-splines with difference penalties on the rows and columns of coefficients, see Eilers & Marx[**?**]). The approach was introduced in [4] for a single event type and we will extend it here to the competing risks setting.

The (u, s)-plane is divided into $n_u \times n_s$ small bins of size h_u and h_s, respectively (rectangles if $h_u \neq h_s$ and squares otherwise), covering the range of observed values of u and s. The bin widths h_u and h_s will be chosen relatively narrow and hence the number of bins n_u and n_s can be relatively large. Within each bin the number of events of type ℓ, denoted by y^ℓ_{jk}, and the total time at risk r_{jk} are determined ($j = 1, \ldots, n_u$; $k = 1, \ldots, n_s$), summed over all individuals $i, i = 1, \ldots, n$. (In the following the superscript ℓ, like in y^ℓ_{jk}, indicates the cause, $\ell = 1, 2$, and not a power.)

The y^ℓ_{jk} can be thought of as realizations of Poisson variates (see [12, 13]) with means

$$\breve{\mu}^\ell_{jk} = r_{jk} \cdot \breve{\lambda}^\ell_{jk} = r_{jk} \cdot \exp\{\breve{\eta}^\ell_{jk}\}, \tag{3}$$

where the $\breve{\lambda}^\ell_{jk}$ represent the cause-specific hazards $\breve{\lambda}_\ell(u, s)$ evaluated at the center of bin (j, k). The $\breve{\eta}^\ell_{jk}$ are the corresponding values of the log-hazard, $\breve{\eta}^\ell_{jk} = \ln \breve{\lambda}^\ell_{jk}$.

Because of the same-sized bins, event counts and at-risk times are on a regular grid and are naturally arranged as $n_u \times n_s$ matrices $\mathbf{Y}_\ell = [y^\ell_{jk}]$ and $\mathbf{R} = [r_{jk}]$. Correspondingly, we denote $\mathbf{M}_\ell = [\breve{\mu}^\ell_{jk}]$ and $\mathbf{E}_\ell = [\breve{\eta}^\ell_{jk}]$, so that equation (3) can be written more concisely in matrix form as

$$\mathbf{Y}_\ell \sim \text{Poisson}(\mathbf{M}_\ell) \qquad \text{with} \qquad \mathbf{M}_\ell = \mathbf{R} \odot \exp\{\mathbf{E}_\ell\}. \tag{4}$$

Here $\odot$ denotes elementwise multiplication.

To obtain smooth cause-specific hazard surfaces we model the log-hazards $\breve{\eta}_\ell(u, s) = \ln \breve{\lambda}_\ell(u, s)$ via tensor products of B-splines. Modeling the log-hazard will automatically produce positive estimates for the cause-specific hazards.

The tensor products are formed from two marginal B-splines bases along the u- and s-axis, with c_u and c_s elements, respectively. We choose cubic B-splines along both axes. The two basis matrices are denoted by $\mathbf{B}_u$, which is $n_u \times c_u$, and by $\mathbf{B}_s$, which is $n_s \times c_s$. The $c_u c_s$ regression coefficients are denoted by α_{fg}^{ℓ} ($f = 1, \ldots, c_u$; $g = 1, \ldots, c_s$), and the log-hazards, evaluated at the bin midpoints, are

$$\breve{\eta}_{jk}^{\ell} = \sum_{f=1}^{c_u} \sum_{g=1}^{c_s} b_{jf}^{u} b_{kg}^{s} \alpha_{fg}^{\ell}. \tag{5}$$

If we arrange the coefficients in the $c_u \times c_s$ matrix $\mathbf{A}_\ell = [\alpha_{fg}^{\ell}]$, then the linear predictor in (5) can be expressed in matrix form as $\mathbf{E}_\ell = \mathbf{B}_u \mathbf{A}_\ell \mathbf{B}_s^{\mathsf{T}}$.

The number of basis functions c_u and c_s, respectively, will be rich enough to allow sufficient flexibility, but difference penalties on the coefficients, both along the rows and the columns of $\mathbf{A}_\ell$, will prevent over-fitting. The penalty on the coefficients is constructed from two matrices $\mathbf{D}_u$ and $\mathbf{D}_s$ that form differences of order d (usually $d = 1$ or $d = 2$) of neigbouring elements in the the columns of a matrix and it is controlled by two smoothing parameters ϱ_u and ϱ_s to allow anisotropic smoothing:

$$\text{pen}(\varrho_u, \varrho_s) = \varrho_u \, ||\mathbf{D}_u \mathbf{A}_\ell||_{\text{F}}^2 + \varrho_s \, ||\mathbf{A}_\ell \mathbf{D}_s^{\mathsf{T}}||_{\text{F}}^2 \tag{6}$$

(The Frobenius norm $||.||_{\text{F}}^2$ is the sum of all squared elements of a matrix.)

The objective function to be minimized is the sum of the Poisson deviance resulting from (3) and the above penalty (6)

$$\text{dev}(\mathbf{M}_\ell; \mathbf{Y}_\ell) + \text{pen}(\varrho_u, \varrho_s) = 2 \sum_{j=1}^{n_u} \sum_{k=1}^{n_s} \left(y_{jk}^{\ell} \ln(y_{jk}^{\ell} / \breve{\mu}_{jk}^{\ell}) - (y_{jk}^{\ell} - \breve{\mu}_{jk}^{\ell}) \right) + \text{pen}(\varrho_u, \varrho_s) \,, \tag{7}$$

which leads to normal equations that can be solved, for given ϱ_u and ϱ_s, in a penalized Poisson IWLS scheme (in compact notation; $\otimes$ denotes the Kronecker product):

$$\left[(\mathbf{B}_s \otimes \mathbf{B}_u)^{\mathsf{T}} \tilde{\mathbf{W}}_\ell (\mathbf{B}_s \otimes \mathbf{B}_u) + \mathbf{P} \right] \tilde{\mathbf{A}}_\ell = (\mathbf{B}_s \otimes \mathbf{B}_u)^{\mathsf{T}} \tilde{\mathbf{W}}_\ell \tilde{\mathbf{z}}_\ell.$$

In the penalty matrix $\mathbf{P} = \rho_u (\mathbf{I}_s \otimes \mathbf{D}_u^{\mathsf{T}} \mathbf{D}_u) + \rho_s (\mathbf{D}_s^{\mathsf{T}} \mathbf{D}_s \otimes \mathbf{I}_u)$ the $\mathbf{I}_u$ and $\mathbf{I}_s$ are identity matrices of appropriate dimension, $\mathbf{W}_\ell$ is a diagonal matrix of weights, $\mathbf{z}_\ell$ is the working variable and the tilde indicates the current value in the iteration.

The smoothing parameters ϱ_u and ϱ_s control the smoothness of the estimated CSHs. Their optimal values are selected by minimizing the AIC (Akaike information criterion), which balances fidelity to the data, as measured by the deviance, and model complexity measured by the effective dimension (ED). This is done by numerically optimizing the AIC of the model as a function of (ϱ_u, ϱ_s). Efficient computation of the model estimates is achieved by employing generalized linear array methods [14], which is possible due to the regular grid of bins underlying $\mathbf{Y}_\ell$, $\mathbf{R}$ and $\mathbf{E}_\ell$.

2.2 Ungrouping final interval by the penalised composite link model

Age information in data on cancer incidence or mortality is often available in age-groups only. Frequently the highest ages are collected in a open-age last interval. Such data grouping is done to prevent identification of the patients [11] or to provide a compact representation of the data. Grouped data do not represent a problem as such, but in some occasions it is desirable to work with data that are disaggregated to a finer resolution. The hazard model with two time scales that was introduced in the previous subsection employs data which are grouped in relatively narrow bins of same size, resulting in data on a regular grid. The gridded structure allows efficient computations by using GLAM algorithms.

The SEER data, which we will analyse in Section 3, provide age at diagnosis in single years up to age 89, but group all observations with higher age at diagnosis in a single last age-group 90+. Time since diagnosis is still given in months for all individuals in this last age-interval, so only one time dimension is affected by coarse grouping. Figure 1, left, illustrates this data format for the number of deaths due to breast cancer (in the subgroup of white women diagnosed with less aggressive cancers and receiving no chemotherapy).

To preserve the gridded data structure we will disaggregate the final age-group using the penalized composite

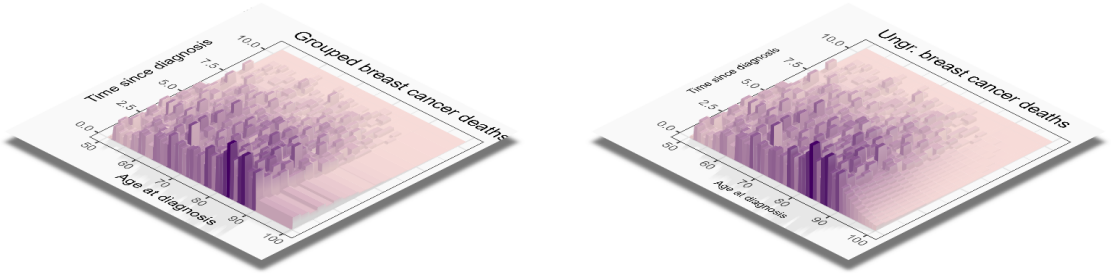

Figure 1: Left: Histogram of deaths due to breast cancer over age at diagnosis (with ages ≥ 90 grouped) and time since diagnosis. Right: Data ungrouped by the PCLM.

link model (PCLM), originally proposed by Eilers [15] and used for ungrouping ages at death in one dimension by Rizzi et al. [16]. Here we use the bivariate extension proposed in [11]. Figure 1, right shows the corresponding ungrouped distribution.

2.3 From cause-specific hazards to cumulative incidence functions

Because of the relationship between the time scales t and s and the fixed time covariate u, it is possible to interpret the cause-specific hazard with two time scales $\breve{\lambda}_\ell(u,s)$ equivalently as a series of one time scale hazards along s that vary smoothly along u values. In other words, each age at diagnosis u has its own CSH over s. This relationship allows us to estimate each of the quantities that derive from the CSHs over the (u,s)-plane and to represent them over the (t,s)-plane without loss of information.

For simplicity, consider first CSHs with one time scale $\lambda_\ell(s)$. The cause-specific cumulative hazard is obtained from $\lambda_\ell(s)$, by integrating over the values of s:

$$\Lambda_\ell(s) = \int_0^s \lambda(v)dv.$$

Over the (u,s)-plane this corresponds to integration over the s time scale only, and no integration over the u scale is required, because u is not a time scale. The cumulated CSH with two time scale is defined as

$$\Lambda_\ell(t,s) = \int_0^s \lambda_\ell(t = u+v, v)\, dv, \tag{8}$$

over the (t,s)-plane or

$$\breve{\Lambda}_\ell(u,s) = \int_0^s \breve{\lambda}_\ell(u,v)\, dv, \tag{9}$$

over the (u,s)-plane.

From the cumulated CSHs of equations 8 and 9, the overall survival function is obtained as

$$S(t,s) = \exp\left\{-\sum_{\ell=1}^{2}\Lambda_\ell(t,s)\right\}, \qquad \text{or} \qquad \breve{S}(u,s) = \exp\left\{-\sum_{\ell=1}^{2}\breve{\Lambda}_\ell(u,s)\right\}, \tag{10}$$

respectively. The overall survival function is the probability of being event-free at (t,s), or equivalently, the probability of being event-free after s time units for a subject that entered the risk set at time u.

The cumulative incidence functions (CIFs) are

$$I_\ell(t,s) = \int_0^s \lambda_\ell(t = u+v, v)\, S(t = u+v, v)\, dv, \quad \text{or, over the } (u,s)\text{-plane} \quad \breve{I}_\ell(u,s) = \int_0^s \breve{\lambda}_\ell(u,v)\, \tilde{S}(u,v)\, dv. \tag{11}$$

The integral in equation 9 has no analytical form, therefore we integrate it numerically, using the rectangle rule.

$$\hat{\dot{\Lambda}}_\ell(\dot{u}, \dot{s}) = \sum_{k=1}^{n_s} \breve{\lambda}_\ell(\dot{u}, \dot{s}_k)\Delta_s \,, \tag{12}$$

Δ_s indicates the distance between two consecutive points $\dot{s_k}$ and $\dot{s_{k-1}}$. The overall survival function is obtained as:

$$\hat{\dot{S}}(\dot{u}, \dot{s}_k) = \exp\left\{ -\sum_{\ell=1}^{2} \hat{\dot{\Lambda}}_\ell(\dot{u}, \dot{s}_k) \right\}, \tag{13}$$

Finally, the cumulative incidence functions are obtained from the CSHs and the overall survival function as:

$$\hat{\dot{I}}_\ell(\dot{u}, \dot{s}) = \sum_{k=1}^{n_s} \left\{ \breve{\lambda}_\ell(\dot{u}, \dot{s}_k)\hat{\dot{S}}(\dot{u}, \dot{s}_k)\Delta_s) \right\} \tag{14}$$

3 Mortality of women with breast cancer

3.1 The SEER data

The Surveillance, Epidemiology and End Results (SEER) program from the National Cancer Institute collects and publishes data on cancer incidence and survival covering about 48% of the population of the USA [17]. In this study we use data from the incidence SEER-17 database (November 2022 submission, [18]), that includes all cancer diagnoses up to 2020. WE extract all cases of malignant breast cancer (primary site codes C50.0-C50.9) diagnosed to women between 2010 and 2015, with maximum follow-up to end of 2019.

The key variables for our analysis are age at diagnosis, time since diagnosis, vital status at the end of the follow-up and cause of death. We additionally consider a variable provided by SEER that combine information on race and ethnicity, the breast subtype and an indicator of whether chemotherapy was performed. We select only women who were 50 years old or older at diagnosis of the cancer. Time since diagnosis is available in the registry as survival months, that is the number of months from diagnosis to either death or end of the follow-up. The SEER registers also provide two variables to distinguish between a death which is attributable to the cancer of the record, and any other causes of death. We use these two variables to perform our competing risks model.

Breast subtypes are recorded in the SEER registers as four distinct categories, which are a combination of the ER status, the PR status and the HER2 status. We decided to distinguish between the Luminal A (HR+ and HER2-) subtype and all other subtypes of breast cancers. Luminal A is the most common and least aggressive subtype of breast cancer and it is linked to greater survival probabilities [19]. The SEER registers also report whether chemotherapy was performed. Finally, we distinguish between white women, black women and women with race/ethnicity other than white or black. Previous research has shown that race is a predictor of survival in women with a breast cancer diagnosis [20], independently of socioeconomic factors and for each stage of the cancer.

3.2 Estimating the cause-specific hazards and cumulative incidence functions

Here we illustrate the results for one of the sub-groups, namely white women with Luminal A subtype who did not receive chemotherapy. Figure 2 shows the cause-specific hazard surface on a $\log_{10}-$scale corresponding to the event 'breast cancer death' for this sub-group. The left panel of the figure shows the same surface in a three-dimensional space. The results presented in Figure 2 indicate that the rate of breast cancer death increases with increasing age at diagnosis, while decreasing with increasing time since diagnosis.

From the estimated CSHs we obtain estimates for the cumulative incidence surfaces by applying the procedure outlined in section 2.3. Figure 3 shows the CIFs for the same sub-group corresponding to the event breast cancer death (left panel) and for all other causes of death (right panel).

References

[1] Putter H, Fiocco M and Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine* 2007; 26(11): 2389–2430. DOI:10.1002/sim.2712. URL `http://dx.doi.org/10.`

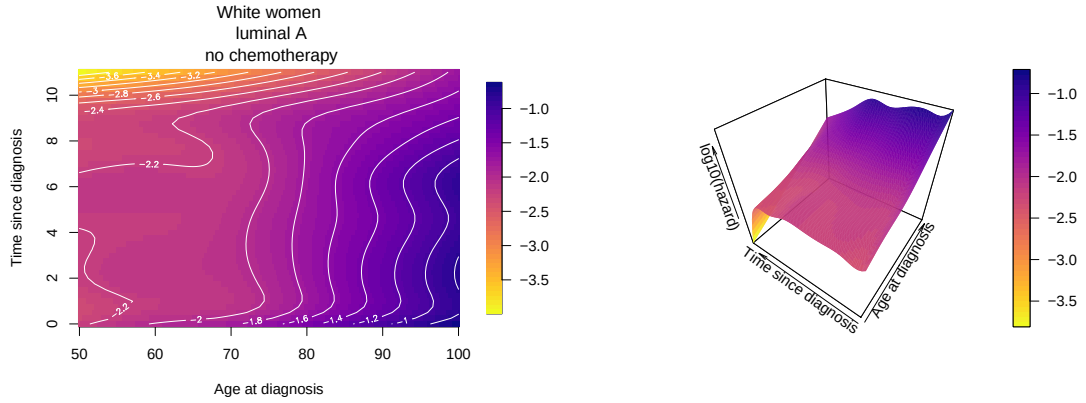

Figure 2: Left: CSH surface of death because of breast cancer for white women with Luminal A subtype and no chemotherapy. Right: the same surface is represented in 3D.

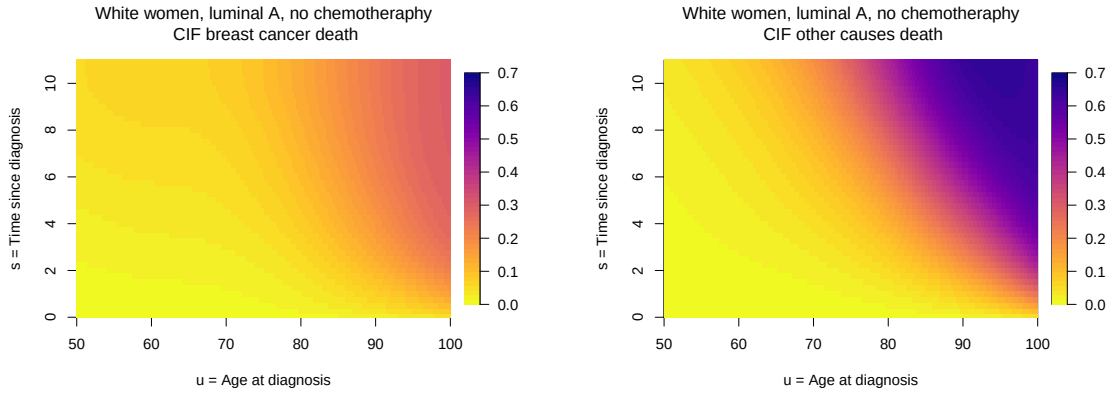

Figure 3: Left: CIF of breast cancer death. Right: CIF of other causes of death. White women with Luminal A subtype and no chemotherapy.

1002/sim.2712.

[2] Berzuini C and Clayton D. Bayesian analysis of survival on multiple time scales. *Statistics in Medicine* 1994; 13(8): 823–838. DOI:10.1002/sim.4780130804. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.4780130804. https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.4780130804.

[3] Skourlis N, Crowther MJ, Andersson TML et al. On the choice of timescale for other cause mortality in a competing risk setting using flexible parametric survival models. *Biometrical Journal* 2022; 64(7): 1161–1177. DOI:10.1002/bimj.202100254.

[4] Carollo A, Eilers PHC, Putter H et al. Smooth hazards with multiple time scales. *arXiv preprint: arXiv:230509342* 2023; http://arxiv.org/abs/2305.09342v1.

[5] Currie ID, Durban M and Eilers PHC. Smoothing and forecasting mortality rates. *Statistical Modelling* 2004; 4(4): 279–298. DOI:10.1191/1471082X04st080oa. URL https://doi.org/10.1191/1471082X04st080oa. https://doi.org/10.1191/1471082X04st080oa.

[6] Arnold M, Morgan E, Rumgay H et al. Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast* 2022; 66: 15–23. DOI:10.1016/j.breast.2022.08.010.

[7] Stapleton SM, Oseni TO, Bababekov YJ et al. Race/Ethnicity and Age Distribution of Breast Cancer Diagnosis in the United States. *JAMA Surgery* 2018; 153(6): 594. DOI:10.1001/jamasurg.2018.0035.

[8] Brandt J, Garne J, Tengrup I et al. Age at diagnosis in relation to survival following breast cancer: a cohort study. *World Journal of Surgical Oncology* 2015; 13(1): 33. DOI:10.1186/s12957-014-0429-x.

[9] Jayasekara H, MacInnis RJ, Chamberlain JA et al. Mortality after breast cancer as a function of time since diagnosis by estrogen receptor status and age at diagnosis. *International Journal of Cancer* 2019; 145(12): 3207–3217. DOI:10.1002/ijc.32214.

[10] National Cancer Institute N. Surveillance, Epidemiology, and End Results Program. Overview of the SEER Program. online, 2023. URL https://seer.cancer.gov/about/overview.html.

[11] Rizzi S, Halekoh U, Thinggaard M et al. How to estimate mortality trends from grouped vital statistics. *International Journal of Epidemiology* 2018; 48(2): 571–582. DOI:10.1093/ije/dyy183.

[12] Holford TR. The analysis of rates and of survivorship using log-linear models. *Biometrics* 1980; : 299–305.

[13] Laird N and Olivier D. Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques. *Journal of the American Statistical Association* 1981; 76(374): 231–240. DOI:10.1080/01621459.1981.10477634.

[14] Currie ID, Durban M and Eilers PHC. Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006; 68(2): 259–280. DOI:10.1111/j.1467-9868.2006.00543.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2006.00543.x. https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2006.00543.x.

[15] Eilers PHC. Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling* 2007; 7(3): 239–254. DOI:10.1177/1471082x0700700302.

[16] Rizzi S, Gampe J and Eilers PHC. Efficient Estimation of Smooth Distributions From Coarsely Grouped Data. *American Journal of Epidemiology* 2015; 182(2): 138–147. DOI:10.1093/aje/kwv020.

[17] National Cancer Institute. Surveillance, Epidemiology, and End Results Program, 2023. URL https://seer.cancer.gov/.

[18] Surveillance, Epidemiology, and End Results (SEER) Program (wwwseercancergov). SEER*Stat Database: Incidence - SEER Research Data, 17 Registries, Nov 2022 Sub (2000-2020) , 2023. - Linked To County Attributes - Time Dependent (1990-2021) Income/Rurality, 1969-2021 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, released April 2023, based on the November 2022 submission.

[19] Howlader N, Cronin KA, Kurian AW et al. Differences in Breast Cancer Survival by Molecular Subtypes in the United States. *Cancer Epidemiology, Biomarkers & Prevention* 2018; 27(6): 619–626. DOI:10.1158/1055-9965.epi-17-0627.

[20] Campbell JB. Breast Cancer-Race, Ethnicity, and Survival: A Literature Review. *Breast Cancer Research and Treatment* 2002; 74(2): 187–192. DOI:10.1023/a:1016178415129.