Hennerfeind, Held:

A Bayesian geoadditive relative survival analysis of registry data on breast cancer mortality

Sonderforschungsbereich 386, Paper 515 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

A Bayesian geoadditive relative survival analysis of registry data on breast cancer mortality

Andrea Hennerfeind

Department of Statistics
Ludwig-Maximilians-University Munich
andreah@stat.uni-muenchen.de

Leonhard Held

Biostatistics Unit
University of Zurich
leonhard.held@ifspm.unizh.ch

Abstract

In this paper we develop a so called relative survival analysis, that is used to model the excess risk of a certain subpopulation relative to the natural mortality risk, i.e. the base risk that is present in the whole population. Such models are typically used in the area of clinical studies, that aim at identifying prognostic factors for disease specific mortality with data on specific causes of death being not available. Our work has been motivated by continuous-time spatially referenced survival data on breast cancer where causes of death are not known. This paper forms an extension of the analyses presented in Sauleau, Hennerfeind, Buemi and Held (2007), where those data are analysed via a geoadditive, semiparametric approach, however without allowance to incorporate natural mortality. The usefulness of this relative survival approach is supported by means of a simulated data set.

Key words: Relative Survival, Bayesian penalized splines, Gaussian Markov Random Fields, MCMC, structured hazard regression, breast cancer

1 Introduction

Many clinical studies aim at identifying prognostic factors for disease specific mortality. However, data on specific causes of death is often not available or not reliable (Percy, Stanek and Gloeckler (1981)) and thus it is not possible to differentiate between cases of

death that are actually related to the disease of interest and those cases of death that are related to other causes that are independent of this disease. Since the composition of patients in a clinical study usually is quite heterogeneous concerning covariates like age (which is the main influencing factor for natural mortality), the natural mortality risk may differ heavily between patients. Thus it might very well be the case that a higher number of deaths is observed with older people although a disease is more likely to be lethal with younger people. In such situations the Cox model is not suitable since therewith it is not possible to distinguish whether a variable such as sex or age has an effect on disease specific mortality, on natural mortality or on both. Consequently this model will deliver effects that represent some mixture of the effects on natural and disease related mortality and may therefore be misleading regarding the identification of prognostic factors. Moreover, comparisons of the results from different population–based prognostic studies are difficult due to differences in the natural mortality of the populations. A remedy to this problem is provided by a relative survival analysis which allows for a correction for the effect of other independent causes of death by using the natural mortality in the underlying population as a reference.

Several models for relative survival analysis in a frequentist setting have been discussed in the literature. Esteve, Benhamou, Croasdale and Raymond (1990) assume that the observed hazard for total mortality is the sum of two hazards, namely the expected, natural mortality hazard and a disease related mortality hazard. Whereas the first component is obtained from external sources the disease related hazard is estimated parametrically assuming a piecewise constant baseline effect and time–constant fixed effects of covariates. This approach was extended by Bolard, Quantin, Esteve, Faivre and Abrahamowicz (2001) and Giorgi et al. (2003) by allowing for time–varying effects, i.e. dropping the proportional hazards assumption. Bolard et al. (2001) consider time–by–covariate interactions originally proposed by Cox (1972) as well as piecewise proportional hazards, developed by Moreau, Le Minor, Myquel, Lellouch (1985) for ordinary survival analysis. The drawbacks of these methods are that temporal variations in the effects of covariates are limited to pre–specified parametric forms of interaction functions and step–functions on pre–specified time intervals, respectively. A more flexible method is proposed by Giorgi et al. (2003) who assume quadratic B–splines with two inner knots for the baseline effect as well as for time–varying effects of covariates. In the Bayesian approach we present here

we extend the model of Esteve et al. (1990) by modelling the disease related hazard with a flexible geoadditive predictor as developed for ordinary survival models in Hennerfeind, Brezger & Fahrmeir (2006), that may include a log–baseline effect, nonlinear effects of continuous covariates and time–varying effects modelled by penalized splines (P–splines), as well as a spatial effect, random effects and the usual fixed effects.

The rest of this article is organized as follows. In Section 2 we describe models, likelihood and priors for unknown functions and parameters. Some comments on the inference via MCMC are given in Section 3. To illustrate our approach we present an application to data on the survival of women suffering from breast cancer in Section 4. Reliability of our approach is verified in Section 5 by means of a simulated data set with known risk profile.

2 Model, likelihood and priors

2.1 Observation model and likelihood

Consider right–censored survival data in usual form, i.e., it is assumed that each individual i in the study has a lifetime T_i (denoting survival time until death of any cause) and a censoring time C_i that are independent random variables. The observed lifetime is then $t_i = \min(T_i, C_i)$, and δ_i denotes the non-censoring indicator given by

$$\delta_i = \begin{cases} 1 & T_i \leq C_i \\ 0 & \text{else} \end{cases}$$

The data are then given by

$$(t_i, \delta_i; a_i, cov_i), \quad i = 1, \ldots, n, \tag{1}$$

where a_i denotes age at diagnosis and cov_i is the vector of covariates (possibly including age as well). Covariates may also be time–dependent, but we restrict discussion to time–constant covariates for simplicity. The same applies to left truncation, which might easily be included, but it is not discussed here for facility of inspection. Following Esteve et al. (1990) we assume that the hazard rate for total mortality $\lambda_i(t, a_i, cov_i) := \lambda_i(t)$ at time t

after diagnosis of an individual i is defined as the following sum of two hazards:

$$\begin{aligned} \lambda_i(t, a_i, cov_i) := \lambda_i(t) &= \lambda_i^e(a_i + t, cov_i^{sub}) + \lambda_i^c(t, cov_i) \\ &= \lambda_i^e(a_i + t, cov_i^{sub}) + \exp\left(\eta_i(t, cov_i)\right) \end{aligned} \tag{2}$$

The first summand $\lambda_i^e(a_i + t, cov_i^{sub})$ represents the expected hazard for natural mortality in a population and is obtained from mortality tables using external sources, i.e. there are no unknown parameters involved here. This component depends only on age at time t after diagnosis (i.e. $a_i + t$) and cov_i^{sub}, a subvector including those covariates in cov_i mortality tables account for (usually sex and period). The second summand $\lambda_i^c(t, cov_i)$ is the disease related mortality hazard rate which is estimated from the data at hand. This component is modelled by a flexible, possibly geoadditive predictor. To simplify notation the dependence on cov_i^{sub} and cov_i, respectively will be suppressed in the following, i.e. we define $\lambda_i^e(a_i + t, cov_i^{sub}) := \lambda_i^e(a_i + t)$ and $\lambda_i^c(t, cov_i) := \lambda_i^c(t)$. Depending on what kind of covariates are given in $cov_i := (z_i, x_i, s_i, v_i)$, the predictor may be composed of the following summands:

$$\eta_i(t, cov_i) := \eta_i(t) = g_0(t) + \sum_{j=1}^{p} g_j(t) z_{ij} + \sum_{j=1}^{q} f_j(x_{ij}) + f_{spat}(s_i) + \boldsymbol{v}_i'\boldsymbol{\gamma} + b_{g_i}, \tag{3}$$

where $g_0(t) = \log\{\lambda_0(t)\}$ is the disease related log–baseline hazard, $g_j(t)$ are time–varying effects of covariates z_j, and $f_j(x_j)$ is the nonlinear effect of a continuous covariate x_j. The function $f_{spat}(s)$ is a (structured) spatial effect, where s, $s = 1, \ldots, S$ is either a spatial index, with $s_i = s$ if subject i is associated with area s, or an exact spatial coordinate $s = (x_s, y_s)$, e.g. for centroids of regions or if exact locations of individuals are known. The vector $\boldsymbol{\gamma}$ is the vector of usual linear fixed effects, and b_g is a subject– or group–specific frailty or random effect, with $b_{gi} = b_g$ if individual i belongs to group g, $g = 1, ..., G$. For $G = n$, we obtain individual–specific frailties, for $G < n$, b_g might be the effect of center g in a multicenter study or the unstructured (uncorrelated random) spatial effect of an area (i.e. $b_g = b_s$), for example. Random slopes could also be introduced, but we omit this here. For identifiability reasons, we center all unknown functions around zero, and include an intercept term in the parametric linear term.

Once more, for a interpretation of equation (2) one may say that the natural mortality hazard λ_i^e covers the basic mortality risk a population is exposed to and the disease related hazard λ_i^c models the excess mortality risk that patients are exposed to beyond the basic

risk due to the disease they suffer from. From a statistical point of view λ_i^e is an additive offset.

Under the assumption about noninformative censoring the likelihood is given by

$$L = \prod_{i=1}^{n} (\lambda_i(t_i))^{\delta_i} \cdot \exp\left(-\int_0^{t_i} \lambda_i(u)du\right)$$

Inserting (2) results in

$$\begin{aligned} L &= \prod_{i=1}^{n} (\lambda_i^e(a_i+t_i) + \lambda_i^c(t_i))^{\delta_i} \exp\left(-\int_0^{t_i} (\lambda_i^e(a_i+u) + \lambda_i^c(u))\, du\right) \\ &= \prod_{i=1}^{n} (\lambda_i^e(a_i+t_i) + \lambda_i^c(t_i))^{\delta_i} \exp\left(-\int_0^{t_i} \lambda_i^c(u)du\right) \exp\left(-\int_0^{t_i} \lambda_i^e(a_i+u)du\right), \quad (4) \end{aligned}$$

where the last factor does not depend on the parameters to be estimated. Hence the following proportionality holds

$$L \propto \prod_{i=1}^{n} (\lambda_i^e(a_i+t_i) + \lambda_i^c(t_i))^{\delta_i} \exp\left(-\int_0^{t_i} \lambda_i^c(u)du\right). \quad (5)$$

This formula only differs from the likelihood of an ordinary survival model of the form

$$\lambda_i(t) = \lambda_i^c(t, cov_i) = \exp\left(\eta_i(t, cov_i)\right),$$

were natural mortality is not accounted for, by the term $\lambda_i^e(a_i + t_i)$.

To obtain a unified and generic notation, we rewrite the observation model in general matrix notation. This is useful for defining priors in the next subsection and for developing posterior analysis in Section 3.

Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_i, \ldots, \eta_n)'$ denote the predictor vector, where $\eta_i := \eta_i(t_i)$ is the value of predictor (3) at the observed lifetime $t_i, i = 1, \ldots, n$. Correspondingly, let $\boldsymbol{g}_j = (g_j(t_1), \ldots, g_j(t_n))'$ denote the vector of evaluations of the functions $g_j(t), j = 0, \ldots, p$, $\boldsymbol{f}_j = (f_j(x_{1j}), \ldots, f_j(x_{nj}))'$ the vector of evaluations of the functions $f_j(x_j), j = 1, \ldots, q$, $\boldsymbol{f}_{spat} = (f_{spat}(s_1), \ldots, f_{spat}(s_n))'$ the vector of spatial effects, and $\boldsymbol{b} = (b_{g_1}, \ldots, b_{g_n})'$ the vector of uncorrelated random effects. Furthermore, let $\tilde{\boldsymbol{g}}_j = (g_j(t_1)z_{1j}, \ldots, g_j(t_n)z_{nj})', j = 1, \ldots, p$.

In the following, we can always express vectors $\boldsymbol{g}_0$, $\tilde{\boldsymbol{g}}_j$, $\boldsymbol{f}_j$, $\boldsymbol{f}_{spat}$ and $\boldsymbol{b}$ as the matrix product of an appropriately defined design matrix $\boldsymbol{Z}$, say, and a (possibly high–dimensional) vector $\boldsymbol{\beta}$ of parameters, e.g. $\tilde{\boldsymbol{g}}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, $\boldsymbol{f}_j = \boldsymbol{Z}_j\boldsymbol{\beta}_j$, etc. Then, after reindexing, we can represent the predictor vector $\boldsymbol{\eta}$ in generic notation as

$$\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_0\boldsymbol{\beta}_0 + \ldots + \boldsymbol{Z}_m\boldsymbol{\beta}_m. \quad (6)$$

2.2 Priors for parameters and functions

The Bayesian model formulation is completed by assumptions about priors for parameters and functions. For fixed effect parameters $\boldsymbol{\gamma}$ in (6) we assume locally constant priors $p(\boldsymbol{\gamma}) \propto const$. A weakly informative normal prior would be another choice. Uncorrelated random effects are assumed to be i.i.d. Gaussian, $b_g \sim N(0, \tau_b^2)$ with unknown variance τ_b^2.

Priors for functions and spatial components are defined by a suitable design matrix $\boldsymbol{Z}_j$, $j = 0, \ldots, m$, and a prior for the parameter vector $\boldsymbol{\beta}_j$. The general form of a prior for $\boldsymbol{\beta}_j$ in (6) is

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\boldsymbol{K}_j\boldsymbol{\beta}_j\right), \tag{7}$$

where $\boldsymbol{K}_j$ is a structure or penalty matrix of $\text{rank}(\boldsymbol{K}_j) = r_j$, shrinking parameters toward zero or penalizing too abrupt jumps between neighboring parameters. For P–splines and intrinsic Gaussian Markov random field (GMRF) priors (Rue and Held (2005)), $\boldsymbol{K}_j$ will be rank deficient, i.e., $r_j < d_j = \dim(\boldsymbol{\beta}_j)$, and the prior is partially improper. The variance τ_j^2 is unknown.

For *unknown functions* $f_j(x_j)$ or $g_j(t)$, we assume Bayesian P–spline priors as in Lang and Brezger (2004). Random walk priors, suggested in Fahrmeir and Lang (2001), may be used as smoothness priors for the baseline effect and time–varying covariate effects in a piecewise exponential model, correspond to the special case of P–splines with degree zero. The basic idea of P–spline regression (Eilers and Marx (1996)) is to approximate a function $f_j(x_j)$ as a linear combination of B–spline basis functions B_m, i.e.

$$f_j(x_j) = \sum_{m=1}^{d_j} \beta_{jm} B_m(x_j). \tag{8}$$

The basis functions B_m are B–splines of degree l defined over a grid of equally spaced knots $x_{min} = \xi_0 < \xi_1 < \ldots < \xi_s = x_{max}$, $d_j = l + s$. The number of knots is moderate, but not too small, to maintain flexibility, but smoothness of the function is encouraged by difference penalties for neighboring coefficients in the sequence $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jd_j})'$. The Bayesian analogue are first or second order random walk smoothness priors

$$\beta_{jm} = \beta_{j,m-1} + u_{jm} \qquad \text{or} \qquad \beta_{jm} = 2\beta_{j,m-1} - \beta_{j,m-2} + u_{jm} \tag{9}$$

with i.i.d. Gaussian errors $u_{jm} \sim N(0, \tau_j^2)$ and diffuse priors $p(\beta_{j1}) \propto const$, or $p(\beta_{j1})$ and $p(\beta_{j2}) \propto const$, for the initial values. A first order random walk penalizes abrupt jumps

$\beta_{jm} - \beta_{j,m-1}$, and a second order random walk penalizes deviations from a linear trend. The amount of smoothness or penalization is controlled by the variance τ_j^2, which acts as a smoothness (hyper–)parameter, with hyperprior defined by (11). The joint prior of the regression parameters $\boldsymbol{\beta}_j$ is Gaussian and can be easily computed as a product of conditional densities defined by (9) as

$$\boldsymbol{\beta_j} \mid \tau_j^2 \propto \tau_j^{-r_j} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta_j}'\mathbf{K_j}\boldsymbol{\beta_j}\right), \tag{10}$$

which is the generic form (7).

The penalty matrix $\boldsymbol{K}_j$ is of the form $\boldsymbol{K}_j = \boldsymbol{D'D}$, where $\boldsymbol{D}$ is a first or second order difference matrix. For second order random walks, for example, $\boldsymbol{D}$ is given by

$$\boldsymbol{D}_{d_j-2\times d_j} = \begin{pmatrix} 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \end{pmatrix}.$$

The matrix $\boldsymbol{K}_j$ has band structure which is very useful for computationally efficient MCMC updating schemes, compare Rue and Held (2005). It has rank $r_j = d_j - 1$ and $r_j = d_j - 2$ for first and second order random walk priors, respectively. The $n \times d_j$ design matrix $\boldsymbol{Z}_j$ consists of the basis functions evaluated at the observations x_{ij}, i.e., $\boldsymbol{Z}_j(i,m) = B_m(x_{ij})$. Priors for the unknown functions $g_j(t)$ are defined in complete analogy as in (8) and (9). The design matrix for time–varying effect terms $\tilde{\boldsymbol{g}}_j$, $j = 1, \ldots, p$ is derived as $\boldsymbol{Z}_j(i,m) = z_{ij}B_m(x_{ij})$.

A common choice for approximating smooth curves are quadratic or cubic B–splines and a second order penalty. This specification is also preferred by Eilers and Marx (1996) and Lang and Brezger (2004) in order to obtain sufficiently smooth results. Computationally, linear splines are simpler. The simplest choice are B–splines of degree zero, i.e. $B_m(x) \equiv 1$ over the m–th interval, and $B_m(x) \equiv 0$ elsewhere. Then the effect is approximated by a piecewise constant function, and the function values follow a random walk model as in Fahrmeir and Lang (2001). This special choice, with time t as covariate, is the easiest way to smooth the baseline in the piecewise exponential model; moreover the integral in the likelihood (4) reduces to a sum. With P–splines of higher degree, however, estimation of smooth baseline effects is improved in terms of MSEs (compare Hennerfeind et al. (2006)).

For the *structured spatial effect* $f_{spat}(s)$ we assume GMRF priors. Two–dimensional tensor product P–spline priors, or Gaussian random field (GRF) priors, common in geostatistics

(kriging) would be another choice (see Hennerfeind et al. (2006)). It depends mainly on the data at hand, which of the different approaches leads to the best fit. For data observed on a discrete lattice or on the level of geographical regions as in our application, GMRFs seem to be most adequate, while surface smoothers as 2d P–splines or kriging may be more natural in situations where exact locations are available.

In the case of *GMRF priors* (compare Rue and Held (2005), Section 3.3.2) we define areas as neighbors if they share a common boundary and assume that the effect of an area s is conditionally Gaussian, with the mean of the effects of neighboring areas as expectation and a variance that is inverse proportional to the number of neighbors of area s. Setting $f_{spat}(s) := \beta_s^{spat}$ we have

$$\beta_s^{spat} \,|\, \beta_{s'}^{spat}, s' \neq s \sim N\left(\frac{1}{N_s}\sum_{s' \in \delta_s} \beta_{s'}^{spat}, \frac{\tau_{spat}^2}{N_s}\right),$$

where N_s is the number of neighbors of area s, and $s' \in \delta_s$ denotes that area s' is a neighbor of area s. The $n \times S$ design matrix $\boldsymbol{Z}_{spat}$ is now a 0/1 indicator matrix. Its value in the i–th row and s–th column is 1 if observation i is located in site or region s, and zero otherwise. The $S \times S$ penalty matrix $\boldsymbol{K}_{spat}$ has the form of an adjacency matrix with $\text{rank}(\boldsymbol{K}_{spat}) = r_{spat} = S - 1$. As for one–dimensional functions the amount of spatial smoothness is controlled by the variance τ_{spat}^2. A generalization to weighted means of neighboring areas is possible but not considered here.

In real data applications we do not know how much of the spatial variation is explained by structured, spatially correlated effects and how much by unstructured, uncorrelated effects. Therefore we may fit an additional (unstructured) area–specific random effect. We recommend to interpret only the sum of the two effects, since identifiability is weak in that case.

We routinely assign inverse Gamma priors $IG(a_j; b_j)$

$$p(\tau_j^2) \propto \frac{1}{(\tau_j^2)^{a_j+1}} \exp\left(-\frac{b_j}{\tau_j^2}\right) \tag{11}$$

to all variances. They are proper for $a_j > 0$, $b_j > 0$, and we use $a_j = b_j = 0.001$ as a standard choice for a weakly informative prior.

The Bayesian model specification is completed by assuming that all priors for parameters are conditionally independent, and that all priors are mutually independent.

3 Markov chain Monte Carlo inference

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', ..., \boldsymbol{\beta}_m')'$ denote the vector of all regression coefficients in the generic notation (6), $\boldsymbol{\gamma}$ the vector of fixed effects, and $\boldsymbol{\tau}^2 = (\tau_0^2, ..., \tau_m^2)$ the vector of all variance components. Full Bayesian inference is based on the entire posterior distribution, which factorizes into the product of the likelihood and the prior:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2 \mid \text{data}) \propto L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)\, p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2).$$

Due to the (conditional) independence assumptions, the prior factorizes into

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2) = \left\{ \prod_{j=0}^{m} p(\boldsymbol{\beta}_j \mid \tau_j^2) p(\tau_j^2) \right\} p(\boldsymbol{\gamma}),$$

where the last factor can be omitted for diffuse fixed effect priors. The likelihood $L(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\tau}^2)$ is given by inserting (3) into (5). Note that the integral does not require integration over the natural mortality hazard $\lambda_i^e(a_i + t)$ (which is fixed anyway), but just over terms of the form

$$I_i = \int_0^{t_i} \exp\left(g_0(u) + \sum_{j=1}^{p} g_j(u) z_{ij} \right) du,$$

where $g_j(t) = \sum \beta_{jm} B_m(t)$. For linear B–splines, the integrals can be solved analytically, but expressions are rather messy and the computational effort is quite high, see Cai, Hyndman and Wand (2002), Appendix. Following their suggestion, we use simple numerical integration in form of the trapezoidal rule for linear B–splines as well as for the commonly used cubic B–splines, where analytical integration is not possible anyway.

Full Bayesian inference via MCMC simulation is based on updating full conditionals of single parameters or blocks of parameters, given the rest of the data. For updating the parameter vectors $\boldsymbol{\beta}_j$, which correspond to the time–independent functions $f_j(x_j)$, as well as spatial effects $\boldsymbol{\beta}^{spat}$, fixed effects $\boldsymbol{\gamma}$ and random effects $\boldsymbol{b}$, we use a slightly modified version of an MH–algorithm based on iteratively weighted least squares (IWLS) proposals, developed for fixed and random effects by Gamerman (1997) and adapted to generalized additive mixed models in Brezger and Lang (2006).

Suppose we want to update $\boldsymbol{\beta}_j$, with current value $\boldsymbol{\beta}_j^c$ of the chain. Then a new value $\boldsymbol{\beta}_j^p$ is proposed by drawing a random vector from a (high–dimensional) multivariate Gaussian proposal distribution $q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)$, which is obtained from a quadratic approximation of the

log–likelihood by a second order Taylor expansion with respect to $\boldsymbol{\beta}_j^c$, in analogy to IWLS iterations in generalized linear models. More precisely, the goal is to approximate the posterior by a Gaussian distribution, obtained by accomplishing *one* IWLS step in every iteration of the sampler. Then, random samples have to be drawn from a high dimensional multivariate Gaussian distribution with precision matrix and mean

$$\boldsymbol{P}_j = \boldsymbol{Z}_j' \boldsymbol{W}(\boldsymbol{\beta}_j^c) \boldsymbol{Z}_j + \frac{1}{\tau_j^2} \boldsymbol{K}_j, \quad \boldsymbol{m}_j = \boldsymbol{P}_j^{-1} \boldsymbol{Z}_j' \boldsymbol{W}(\boldsymbol{\beta}_j^c)(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{\eta}}).$$

where $\tilde{\eta}_i = \eta_i(t_i) - f_j(x_{ij})$, $\boldsymbol{W}(\boldsymbol{\beta}_j^c) = diag(w_1, \ldots, w_n)$ is the weight matrix for IWLS with weights

$$w_i \;\; = \;\; \Lambda_i^c(t_i) - \frac{\lambda_i^e(a_i + t_i)\lambda_i^c(t_i)\delta_i}{\lambda_i(t_i)^2}$$

obtained form the current state $\boldsymbol{\beta}_j^c$ and with $\Lambda_i^c(t_i) = \int_0^{t_i} \lambda_i^c(u)du$. The working observations $\tilde{y}_i$ are given by

$$\tilde{y}_i = \eta_i(t_i) + \frac{\delta_i \lambda_i^c(t_i)/\lambda_i(t_i) - \Lambda_i^c(t_i)}{w_i}.$$

See Hennerfeind (2006), Appendix A2, for a detailed derivation of those quantities. The proposed vector $\boldsymbol{\beta}_j^p$ is accepted as the new state of the chain with probability

$$\alpha(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p) = \min\left(1, \frac{p(\boldsymbol{\beta}_j^p \mid \cdot) q(\boldsymbol{\beta}_j^p, \boldsymbol{\beta}_j^c)}{p(\boldsymbol{\beta}_j^c \mid \cdot) q(\boldsymbol{\beta}_j^c, \boldsymbol{\beta}_j^p)}\right)$$

where $p(\boldsymbol{\beta}_j \mid \cdot)$ is the full conditional for $\boldsymbol{\beta}_j$ (i.e. the conditional distribution of $\boldsymbol{\beta}_j$ given all other parameters and the data).

4 Application

We illustrate our method by an application to data on breast cancer that was gathered in the years from 1988 to 2002 by a cancer registry that covers the Haut–Rhin 'department' which is located in the north-east of France, adjacent to Germany and Switzerland. This department has 3525 km^2 and 707555 inhabitants (in 1999) and is partitioned into 377 municipalities. The largest distance between the centroids of two municipalities is about 95 kms. The data set contains 3726 cases of breast cancer diagnosed between January the 1st 1988 and January the 1st 1998. There were 1235 ($\approx 33\%$) deaths observed whereas the causes of death are unknown. Observed lifetimes are given in days and range from 0 to 14 years, with a median of 6.4 years. Covariates are age at time of diagnosis (ranging

from 20.6 years to 87.1 years), date of diagnosis (ranging from 1988.0 (i.e. 01.01.1988) to 1998.0), area of residence (one of 377 municipalities) and number of metastases at the date of diagnosis (no metastasis, one metastasis or more than one metastasis). This is part of a data set that has been analyzed via ordinary survival analysis in Sauleau et al. (2007).

For comparison only we analyze the data with an ordinary survival model although this model does not account for natural mortality and is thus not appropriate to the data at hand where causes of death are not available. Generally the specification of the hazard rate is given by

$$\begin{aligned} \lambda_i(t, cov_i) &= \exp(\eta_i(t, cov_i)) \\ cov_i &= (a_i, p_i, s_i, \texttt{meta1}_i, \texttt{meta2}_i), \end{aligned} \tag{12}$$

where t is time since diagnosis and cov_i is the vector of covariates with a_i denoting the age of patient i at date of diagnosis p_i (period), s_i denoting the municipality patient i resides in and the dummy–coded covariates $\texttt{meta1}_i$ and $\texttt{meta2}_i$ denoting, whether patient i has one metastasis and more than one metastasis, respectively.

A relative survival analysis should be more suitable and deliver better results. Therefore we alternatively assume a composed hazard rate of the following structure

$$\begin{aligned} \lambda_i(t, cov_i) &= \lambda_i^e(a_i + t, p_i + t) + \exp(\eta_i(t, cov_i)) \\ cov_i &= (a_i, p_i, s_i, \texttt{meta1}_i, \texttt{meta2}_i), \end{aligned} \tag{13}$$

where $\lambda_i^e(a_i + t, p_i + t)$ is the natural mortality rate of women of age $a_i + t$ at date $p_i + t$ as recorded in mortality tables for the Haut–Rhin department. The second term $\lambda_i^c = \exp(\eta_i(t, cov_i)$ represents the disease related hazard rate and is modelled in the same way as the hazard rate in (12).

A hierarchy of models is analyzed with both approaches and compared via the Deviance Information Criterion (DIC) developed in Spiegelhalter, Best, Carlin and van der Linde (2002). It is given as

$$DIC = D(\overline{\boldsymbol{\theta}}) + 2p_D = \overline{D(\boldsymbol{\theta})} + p_D,$$

where $\boldsymbol{\theta}$ is the vector of parameters, $D(\overline{\boldsymbol{\theta}})$ is the deviance of the model evaluated at the posterior mean estimate $\overline{\boldsymbol{\theta}}$, $\overline{D(\boldsymbol{\theta})}$ is the posterior mean of the deviance and $p_D = \overline{D(\boldsymbol{\theta})} - D(\overline{\boldsymbol{\theta}})$ is the effective number of parameters. Since it is at least unclear, how the

saturated model should be defined in the case of survival data when the baseline hazard and other nonparametric functions are parameters of interest, we use the unstandardized deviance $D(\boldsymbol{\theta}) = -2 \cdot$log–likelihood instead of the saturated deviance. Note however, that the proportionality in (5) was used for the calculation of the DIC in the case of a relative survival model and the DIC resulting from a relative survival model is not to be compared to the DIC resulting from an ordinary survival model.

Whilst a log–baseline effect $g_0(t)$ modelled by a cubic P–spline prior with 20 knots is included in any model, covariate effects are only included gradually. Effects $f_a(a_i)$ and $f_p(p_i)$ of continuous covariates are modelled by cubic P–splines with 20 knots. An unstructured (random) spatial effect b_{s_i} is included additionally or alternatively in some of the models. Table 1 gives values for fit and complexity of a selected number of models according to the two components of the deviance information criterion. Model I, which contains a structured spatial effect modelled by a GMRF–prior, the effect of the number of metastases and the effect of age, yields a DIC of 9308 for the ordinary survival model with hazard rate (12) and 9249 for the relative survival model. Leaving out one or more of these effects leads to a larger DIC. As Table 1 shows the DIC is slightly reduced by the additional inclusion of a period effect. Models III and IV are versions of model II where the spatial effect is modelled by an unstructured (random) effect b_s and the sum of a structured and an unstructured effect, respectively. However, those models will not be discussed here since they do not lead to an improvement in terms of DIC.

Figure 1 displays the estimated nonparametric effects of model II with predictor

$$\eta_i = g_0(t) + f_a(a_i) + f_p(p_i) + f_{spat}(s_i) + \gamma_1 \texttt{meta}_1 + \gamma_2 \texttt{meta}_2.$$

All unknown functions are centered around zero, and an intercept term is included in the parametric linear term for identifiability reasons. For plotting, the estimated effects of age a_i and period p_i are all centered at the observed values, i.e.

$$\sum_{i=1}^{3726} \hat{f}_a(a_i) = \sum_{i=1}^{3726} \hat{f}_p(p_i) = 0,$$

while the intercept is added to the log–baseline effects. Hence it can be derived from Figure 1(a) and (b) that the estimated global risk level is higher with the ordinary survival model (since the log–baseline effect resulting from an ordinary survival analysis exceeds the log–baseline effect resulting from a relative survival analysis). This results from the fact that

the ordinary survival analysis delivers an estimation of the risk of dying of any cause, whereas only the disease related excess mortality risk of breast cancer patients is estimated by means of a relative survival analysis, where the natural mortality risk is accounted for separately. Panels (a) and (b) further reveal that the ordinary survival analysis yields a fairly constant log–baseline effect $g_0(t)$, whereas a relative survival analysis results in an effect, that is increasing in the first two years and decreasing in the time between the third and the 11th year after diagnosis. Presumably the decrease in risk is not reflected in panel (a) as not accounting for natural mortality that is increasing with time after diagnosis (since patients are aging) might lead to a neutralization. The estimated effects of age at time of diagnosis exhibit an u–shaped risk profile and are displayed in panels (c) and (d). While an ordinary survival analysis yields an increased risk for patients diagnosed with breast cancer in their younger days, but a still much higher risk for those women diagnosed at an age of more than 70 years, a relative survival analysis suggests that women diseased in early life have the greatest risk. This result is in accordance with the fact that cancers are often more aggressive with younger people. The differences between the two approaches were to be expected since older women have a higher natural mortality risk that is not accounted for separately with the ordinary, but only with the relative survival analysis.

As displayed in panels (e) and (f) both approaches yield a higher risk for patients that were diagnosed with breast cancer in earlier periods. This effect might be explained by medical progress. Figures 2 (a) and (b) display the values of the structured spatial effect in each municipality. The two approaches yield a similar spatial pattern, but it is more pronounced with the relative survival analysis. The risk seems to be higher in the south of the region. For a nominal level of 95% none of these effects has strictly positive or negative credible intervals, i.e for none of the regions more than 95% of the posterior samples are either all positive or all negative. However, a couple of regions exhibit strictly positive or strictly negative credible intervals for a nominal level of 80%, such as some regions in the north–east that have a lower risk (Figures 2(c) and (d)). The estimated parameters $\hat{\gamma}_1$ and $\hat{\gamma}_2$ for the fixed effects of `meta1` and `meta2` are greater with the relative survival approach. In detail the results are as follows:

	Model	ordinary survival			relative survival		
		$D(\bar{\theta})$	p_D	DIC	$D(\bar{\theta})$	p_D	DIC
I	$g_0(t) + f(a) + f_{spat}(s) + \texttt{meta}$	9268	20	9308	9208	20	9249
II	$g_0(t) + f(a) + f(p) + f_{spat}(s) + \texttt{meta}$	9259	24	9307	9200	23	9246
III	$g_0(t) + f(a) + f(p) + b_s + \texttt{meta}$	9264	24	9312	9205	24	9253
IV	$g_0(t) + f(a) + f(p) + f_{spat}(s) + b_s + \texttt{meta}$	9250	29	9308	9192	28	9248
V	$g_0(t) + f(a) + f(p) + f_{spat}(s) + g(t) * \texttt{meta}$	9239	28	9296	9187	27	9241

Table 1: Deviance, effective number of parameters p_D and DIC for some of the models we compare.

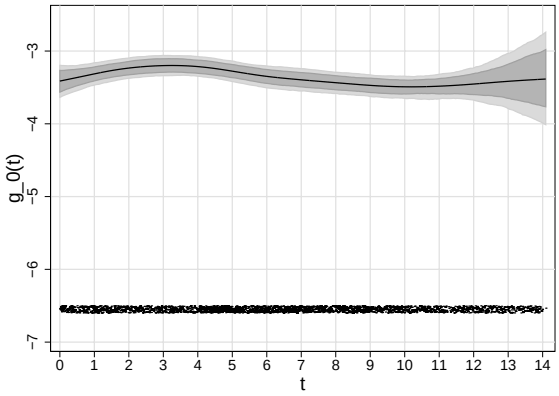
	ordinary	relative
$\hat{\gamma}_1$	0.66 (st.dev. 0.06)	0.96 (std.dev. 0.08)
$\hat{\gamma}_2$	2.23 (st.dev. 0.14)	2.74 (std.dev. 0.15)

meaning that compared to patients with no metastases the hazard rate is about 1.9 (9.3) and 2.6 (15.5) times higher for patients with one (more than one) metastasis, respectively.

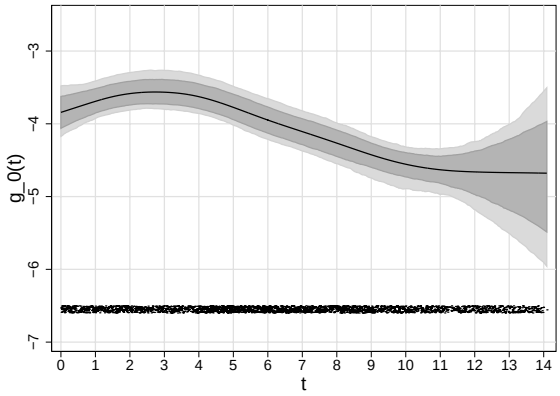
In praxis the proportional hazards assumption does often not hold, and Sauleau et al. (2007) have shown that an ordinary survival model actually gives clear hints for a violation of the proportional hazards assumption. For this reason the number of metastases is included as a covariate with time–varying effect in model II, i.e. the disease–related log–hazard of model V is

$$\lambda_i^c = \exp\left(g_0(t) + \texttt{meta1}_i \cdot g_1(t) + \texttt{meta2}_i \cdot g_2(t) + f_{age}(a_i) + f_p(p_i) + f_{spat}(s_i)\right).$$

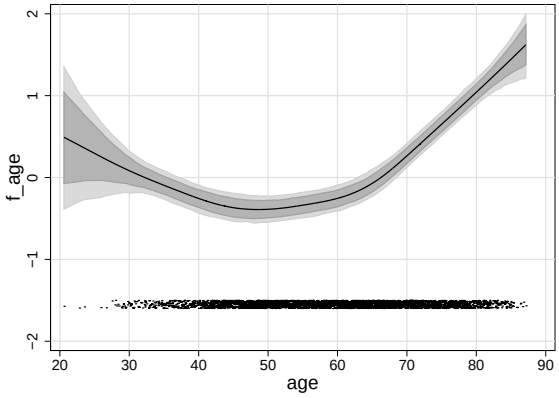
Here $g_0(t)$ is the log–baseline effect for patients without metastases, $g_0(t) + g_1(t)$ corresponds to the log–baseline for patients with one metastasis and $g_0(t) + g_2(t)$ for patients with more than one metastasis. The time–dependent functions $g_k(t), k = 0, 1, 2$ are modelled with cubic P–spline priors with 20 knots. As displayed in Table 1 the DIC is reduced by allowing for a temporal variation in the effect of the number of metastases. The three log–baseline effects are plotted in Figure 3 and reveal that the differences in risk between the patient groups seem to diminish with time after diagnosis. The log–baseline effect for patients with more than one metastasis even crosses the other curves, but this result must not be over–interpreted since there are only 70 patients with more than one metastasis in the study. The remaining estimated effects of model V resemble the results of model II and are not shown for this reason.



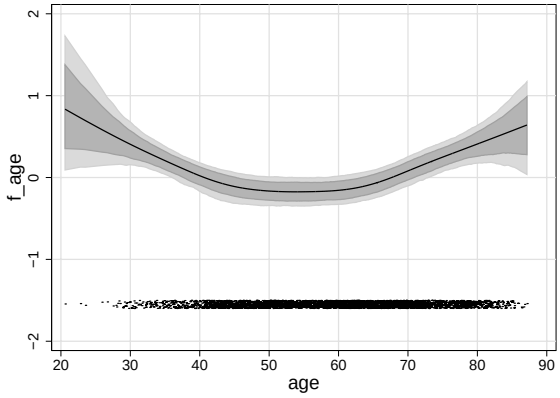
(a) log–baseline effect $g_0(t)$



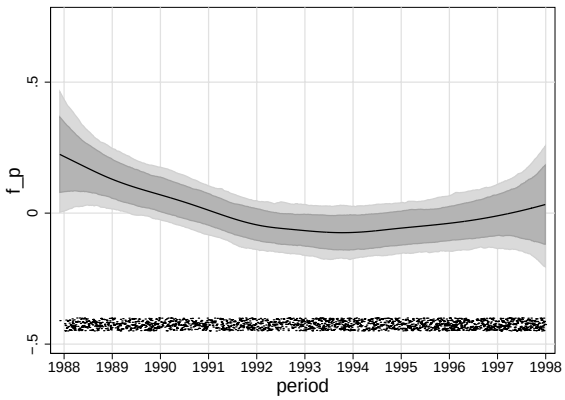
(b) log–baseline effect $g_0(t)$



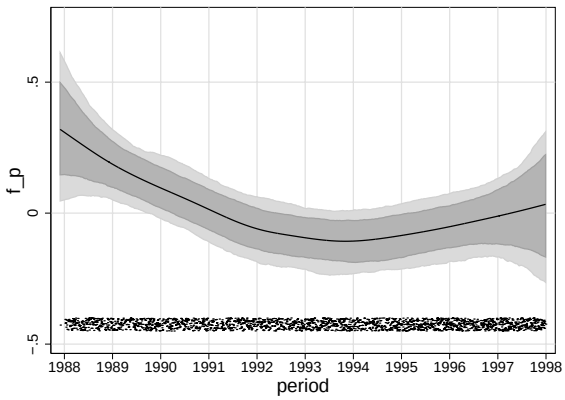
(c) centered effect of age



(d) centered effect of age

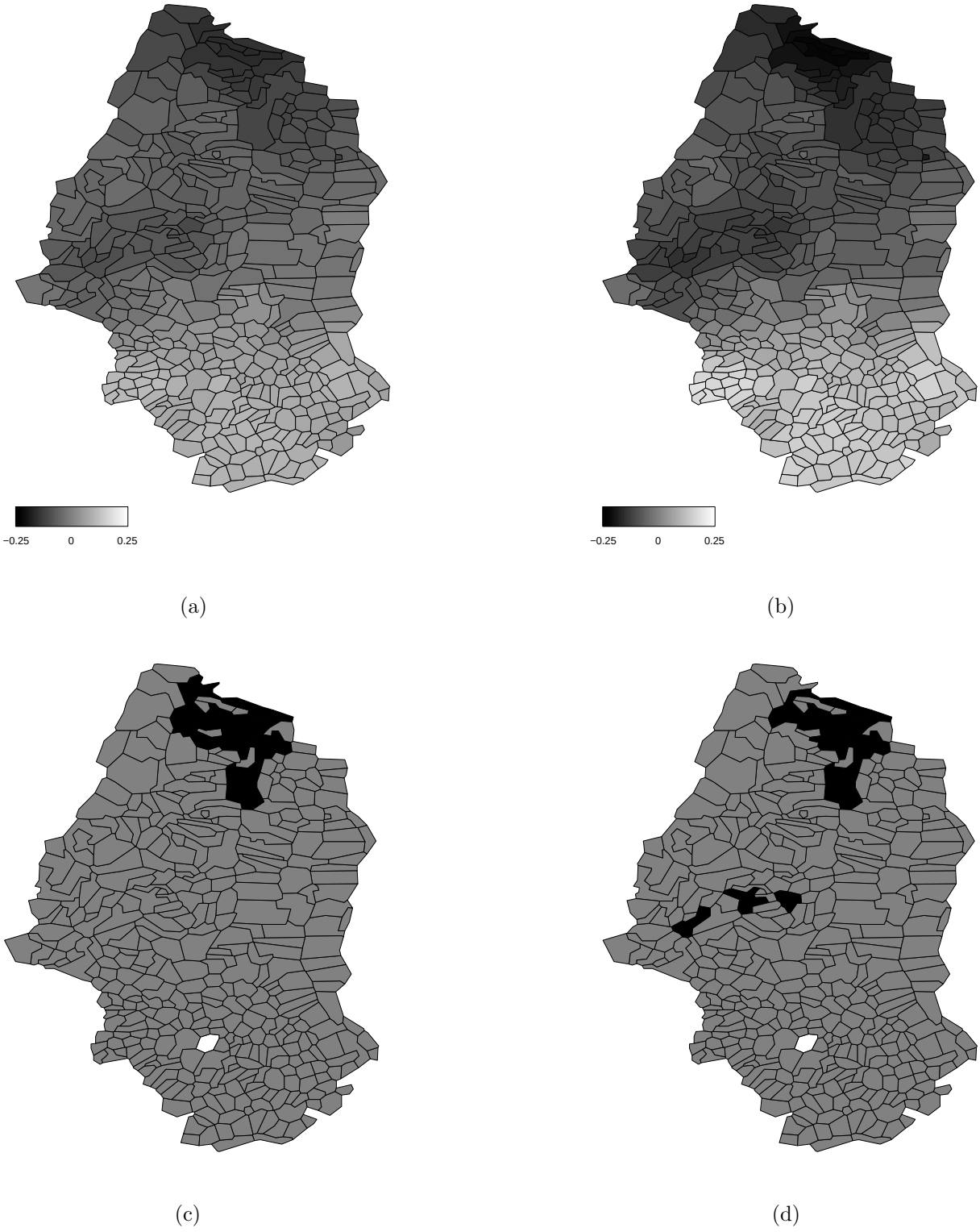


(e) centered effect of period



(f) centered effect of period

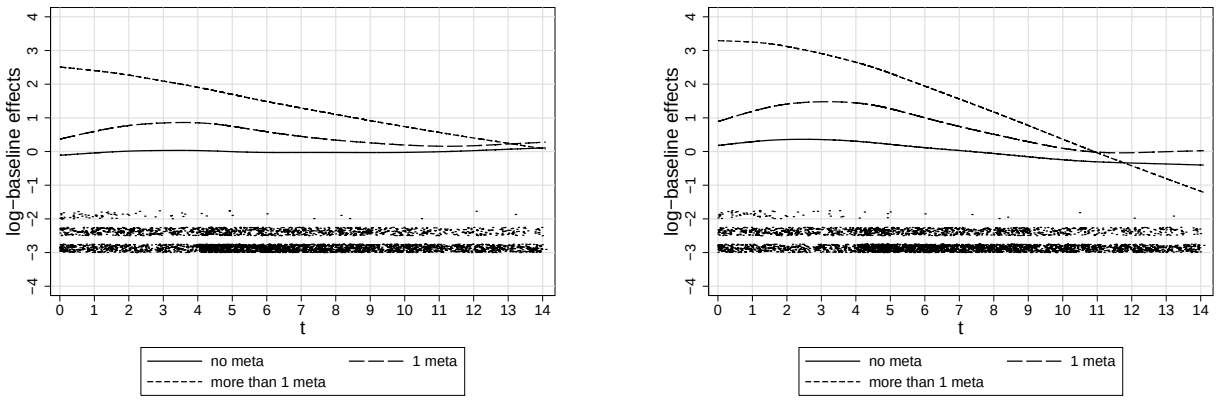
Figure 1: Model II: Posterior means and pointwise 80% and 95% confidence intervals for the baseline effect including the intercept term (a,b), the centered effect of age (c,d) and the centered effect of period (e,f). Figures b,d and f result from a relative survival analysis. Black dots in the lower part of each figure mark observed lifetimes, ages and periods, respectively.



(a) (b)

(c) (d)

Figure 2: Model II: posterior means of the structured spatial effect (a, b) and posterior probabilities for a nominal level of 80% (c, d), where black denotes regions with strictly negative credible intervals and white denotes regions with strictly positive credible intervals. Remaining gray areas in c) and d) exhibit neither strictly positive nor strictly negative credible intervals. Panels b) and d) result from a relative survival analysis.



(a) ordinary survival

(b) relative survival

Figure 3: Model V: posterior means of the log–baseline effects for patients with no metastases, one metastasis and more than one metastasis (dots in the lowest, middle and highest row mark observed lifetimes of patients with no metastases, one metastasis and more than one metastasis, respectively)

5 Simulation

To verify the reliability of our relative survival model and to show that a model that does not account for natural mortality can indeed be misleading concerning the effects of covariates in such cases where data on specific causes of death is not available, we simulate an appropriate data set with known risk profile. Survival times are generated according to a hazard rate that is the sum of a natural hazard rate and a disease related hazard rate. This data set is then analyzed with an ordinary survival model and with a relative survival model like in (2) and the results are compared subsequently.

As for the data generation we simulate survival times based on the covariates of our real breast cancer data set, using known specifications for the baseline and the covariate effects that resemble the effects estimated by the relative survival analysis of the real data set. However, for the sake of simplification we neither consider a spatial effect nor a period effect. We first simulate survival times for each subject and the censoring is done in a second step. In detail, survival times $T_i, i = 1, \ldots, 3726$, are generated according to the following hazard rate model

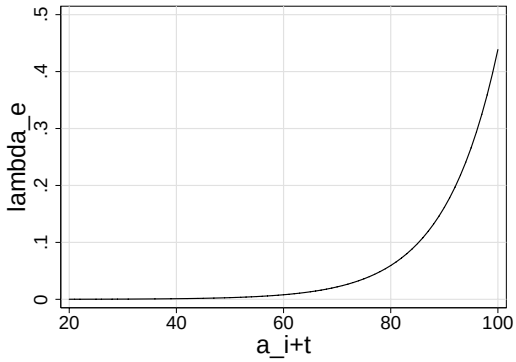
$$\begin{aligned}\lambda_i(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) &= \lambda_i^e(a_i + t) + \lambda_i^c(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) \\ &= \lambda_i^e(a_i + t) + \exp(g_0(t) + f_{age}(a_i) + \gamma_1 \texttt{meta1}_i + \gamma_2 \texttt{meta2}_i),\end{aligned}$$

where the natural hazard rate λ_i^e is chosen in order to resemble the natural mortality rates used with the application, but only depends on $a_i + t$, which is the age of individual i at time t after diagnosis. In our application natural mortality also depends on calendar time, but we did not consider this here. As illustrated in Figure 4(a) the natural hazard rate is increasing exponentially with age at time t after diagnosis. The disease related hazard rate λ_i^c depends on time t after diagnosis, the age at time of diagnosis a_i, and the two binary covariates $\texttt{meta1}_i$ and $\texttt{meta2}_i$, which indicate whether an individual i has one and more than one metastasis, respectively. As displayed in Figure 4(b) the disease related log–baseline $g_0(t)$ is increasing in the first 2.5 years after diagnosis, decreasing in the time span between 2.5 and 12 years and staying constant afterwards. In contrast to the natural mortality risk, the effect of age on the disease related risk is u–shaped and highest with patients diseased in early life, whereas it is less increased with the initially oldest patients in the study, who are diagnosed with breast cancer at the age of 87 (Figure 4(c)). Finally the disease related log–hazard is increased by $\gamma_1 = 0.95$ and $\gamma_2 = 2.75$ for individuals with one metastasis ($\texttt{meta1}_i = 1$) and more than one metastasis ($\texttt{meta2}_i = 1$), respectively. Since the data used in our application were only gathered until the year 2002 we consider all survival times exceeding the year 2002 as censored, i.e. observed survival times are given by $t_i = \min(T_i, 2002.0 - p_i)$ with p_i denoting the exact date of diagnosis observed in the real data set. This mechanism results in a censoring rate of approximately 60% (compared to approximately 67% with the real data set).

The data set generated in this way is initially analyzed with an ordinary survival model that does not distinguish between natural mortality and disease related mortality. More precisely we wrongly assume a hazard rate as follows:

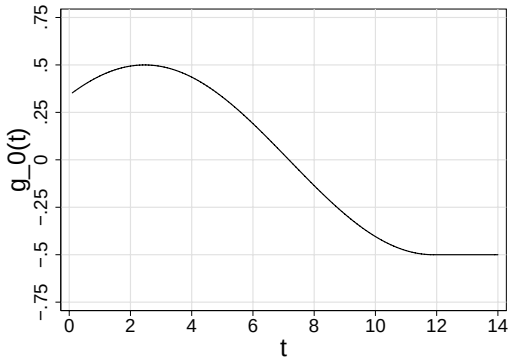
$$\lambda_i(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) = \exp(g_0(t) + f_{age}(a_i) + \gamma_1 \texttt{meta1}_i + \gamma_2 \texttt{meta2}_i),$$

where the log–baseline $g_0(t)$ and the age–effect f_{age} are modelled as cubic P–splines with 20 knots (with second order random walk smoothness priors and $IG(0.001, 0.001)$ priors for the variance components) and γ_1 and γ_2 are fixed effects with diffuse priors. Expectedly the estimated log–baseline and the effect of age do not reflect the true disease related



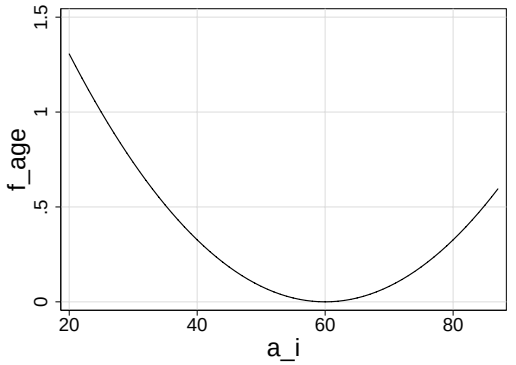
a) natural hazard rate against age

$$\lambda_i^e(a_i + t) = \exp\left((a_i + t - 30)/10\right)/2500$$



b) disease related log–baseline effect

$$g_0(t) = \begin{cases} 0.5 \cdot \sin(t/3 + 0.75) - 4 & , t \leq 4.5\pi - 2.25 \\ -4.5 & , t > 4.5\pi - 2.25 \end{cases}$$



c) disease related effect of age at time of diagnosis

$$f_{age}(a_i) = ((a_i - 60)/35)^2$$

Figure 4: Simulation: specifications for the natural hazard rate, the disease related log–baseline effect and the disease related effect of age at time of diagnosis

effects but rather present a mixture of the two effects on natural mortality and disease related mortality. The estimated log–baseline effect is increasing in the first years after diagnosis, but the subsequent decline is less steep than with the true log–baseline effect (Figure 5(a)). While the disease related log–baseline is decreasing between the 2.5th and 12th year after diagnosis, the natural mortality risk of each single patient is increasing with time (since people are getting older) and these two effects seem to kind of balance. As can be seen from Figure 5(c) the ordinary survival model underestimates the risk for women diagnosed with breast cancer in early years and overestimates the risk of women diseased at an old age. Again, this high risk for older people results from the increasing natural mortality risk that is not accounted for separately. Finally also the fixed effects of the covariates `meta1` and `meta2` are not estimated correctly, but are rather underestimated by $\hat{\gamma_1} = 0.68$ and $\hat{\gamma_2} = 2.33$ (with standard deviations of 0.05 and 0.13, respectively). This underestimation is due to the fact that only a part of the cases of death (namely those cases that are related to the disease) are in association with the number of metastases, whereas the ordinary survival analysis estimates the average influence based on all cases of death.

Now we re–analyze the generated data set with a relative survival model as described in (2). That is we assume a hazard rate as follows:

$$\begin{aligned}\lambda_i(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) &= \lambda_i^e(a_i + t) + \lambda_i^c(t, a_i, \texttt{meta1}_i, \texttt{meta2}_i) \\ &= \frac{\exp\left(\frac{a_i+t-30}{10}\right)}{2500} + \exp(g_0(t) + f_{age}(a_i) + \gamma_1 \texttt{meta1}_i + \gamma_2 \texttt{meta2}_i),\end{aligned}$$

where the disease related hazard rate λ_i^c is modelled as the total hazard rate λ_i was modelled before. However, the total hazard is now amended by the known natural mortality rate λ_i^e in order to account for cases of death that are not related to the disease of interest. As displayed in Figures 5(b) and (d) the true disease related log–baseline and the effect of age are now estimated quite satisfactorily, even though the effect of age is a bit too flat which might be due to the very small number of young patients. Also the fixed effects of $\texttt{meta1}_i$ and $\texttt{meta2}_i$ are estimated quite well with $\hat{\gamma_1} = 0.98$ and $\hat{\gamma_2} = 2.79$ (with standard deviations of 0.07 and 0.15, respectively).

6 Conclusion

In summary it can be ascertained that the simulation supports the usefulness of the relative survival approach since it yields results that are highly comparable to those of our application. As the simulation has shown, a model that does not account for natural mortality is not suitable for the identification of prognostic factors for disease specific mortality in cases where data on causes of death is not available since effects of covariates on natural mortality and effects on disease specific mortality intermix and can not be separated easily ex post.

Acknowledgements: The data analyzed in this article have been kindly provided by Erik–Andre Sauleau. The authors would like to thank Ludwig Fahrmeir and Erik–Andre Sauleau for helpful discussions. Financial support of the German Science Foundation DFG, Sonderforschungsbereich 386 "Statistische Analyse Diskreter Strukturen" is gratefully acknowledged.

References

Bolard, P., Quantin, C., Esteve, J., Faivre, J., and Abrahamowicz, M. (2001), "Modelling time–dependent hazard ratios in relative survival: Application to colon cancer," *Journal of Clinical Epidemiology*, **54**, 986–996.

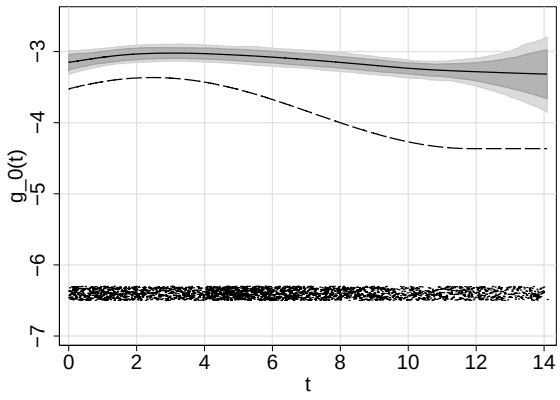
Brezger, A., and Lang, S. (2006), "Generalized structured additive regression based on Bayesian P–splines," *Computational Statistics and Data Analysis*, 50, 967–991.

Cai, T., Hyndman, R., and Wand, M. (2002), "Mixed model-based hazard estimation," *Journal of Computational and Graphical Statistics*, 11, 784–798.

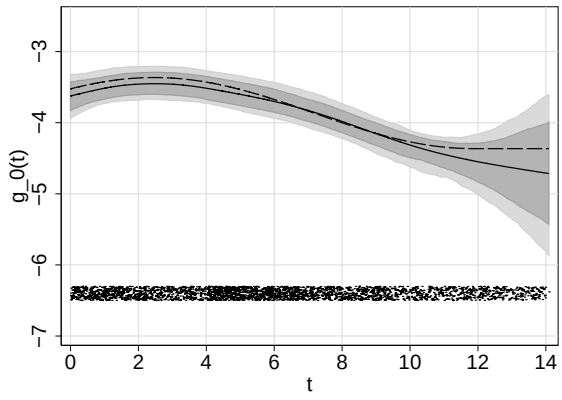
Cox, D.R. (1972), "Regression models and life tables," *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Eilers, P.H.C., and Marx, B.D. (1996), "Flexible smoothing using B-splines and penalized likelihood" (with comments and rejoinder), *Statistical Science*, 11 (2), 89–121.

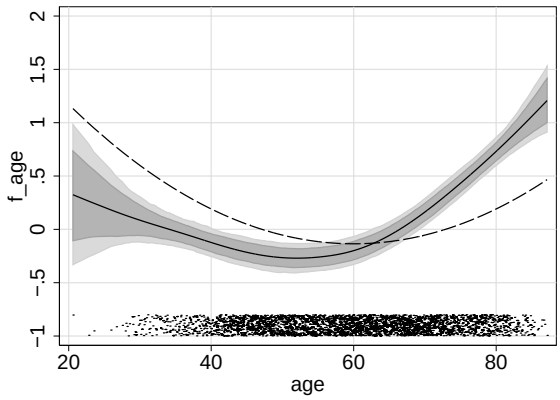
ordinary survival **relative survival**



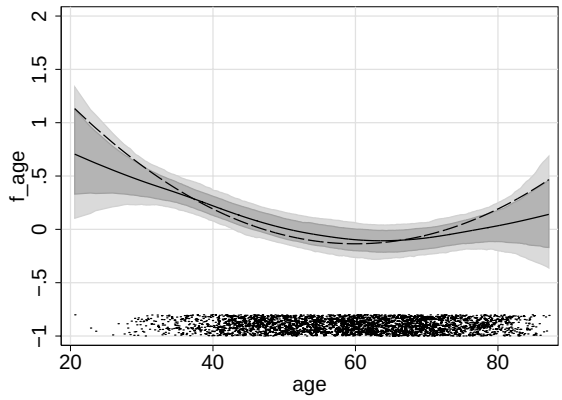
(a) log–baseline effect $g_0(t)$



(b) log–baseline effect $g_0(t)$



(c) centered effect of age



(d) centered effect of age

Figure 5: Simulation: posterior means (solid line) together with pointwise 80% and 95% confidence intervals and true disease related effects (dashed lines) for the log–baseline effect including the intercept term (a,b) and the centered effect of age (c,d). Figures b and d result from a relative survival analysis.

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

Fahrmeir, L., and Lang, S. (2001), "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors," *Journal of the Royal Statistical Society*, Ser. C, 50, 201–220.

Gamerman, D. (1997), "Efficient Sampling from the Posterior Distribution in Generalized Linear Models," *Statistics and Computing*, 7, 57–68.

Giorgi, R., Abrahamowicz, M., Quantin, C., Bolard, P., Esteve, J., Gouvernet, J. and Faivre, J. (2003), "A relative survival regression model using B–spline functions to model non–proportional hazards," *Statistics in Medicine*, **22**, 2767–2784.

Hennerfeind, A. (2006), "Bayesian nonparametric regression for survival and event history data," phD-thesis, LMU Munich, *Dr. Hut-Verlag*.

Hennerfeind, A., Brezger, A., Fahrmeir, L. (2006), "Geoadditive survival models," *Journal of the American Statistical Association*, **101**, 1065–1075.

Lang, S., and Brezger, A. (2004), "Bayesian P–splines," *Journal of Computational and Graphical Statistics*, 13, 183–212.

Moreau, T., Le Minor, M., Myquel, P., Lellouch, J. (1985), "Estimation of the Hazards Ratio in Two Grouped Samples," *Biometrics*, **41**, 245–252.

Percy, C.L., Stanek, E., and Gloeckler, L. (1981), "Accuracy of cancer death certificates and its effect on cancer mortality statistics," *American Journal of Public Health*, **71**, 242–250.

Rue, H., and Held, L. (2005), "Gaussian Markov Random Fields; Theory and Applications", CRC Press, Chapman and Hall.

Sauleau, E.-A., Hennerfeind, A., Buemi, A., and Held, L. (2007), "Age, period and cohort effects in Bayesian smoothing of spatial cancer survival with geoadditive models," *Statistics in Medicine*, to appear.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002), "Bayesian measures of model complexity and fit" (with discussion and rejoinder), *Journal of the Royal Statistical Society*, Ser. B, 64, 583–639.