Received: 16 April 2021 | Accepted: 6 April 2022

DOI: 10.1111/rssc.12566

ORIGINAL ARTICLE

A unifying framework for flexible excess hazard modelling with applications in cancer epidemiology

Alessia Eletti[1] | **Giampiero Marra**[1] | **Manuela Quaresma**[2] | **Rosalba Radice**[3] | **Francisco Javier Rubio**[1]

[1]Department of Statistical Science, University College London, London, UK

[2]Department of Non-Communicable Disease Epidemiology, London School of Hygiene & Tropical Medicine, London, UK

[3]Faculty of Actuarial Science and Insurance, Bayes Business School, City, University of London, London, UK

Correspondence
Alessia Eletti, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
Email: alessia.eletti.19@ucl.ac.uk

Abstract

Excess hazard modelling is one of the main tools in population-based cancer survival research. Indeed, this setting allows for direct modelling of the survival due to cancer even in the absence of reliable information on the cause of death, which is common in population-based cancer epidemiology studies. We propose a unifying link-based additive modelling framework for the excess hazard that allows for the inclusion of many types of covariate effects, including spatial and time-dependent effects, using any type of smoother, such as thin plate, cubic splines, tensor products and Markov random fields. In addition, this framework accounts for all types of censoring as well as left truncation. Estimation is conducted by using an efficient and stable penalized likelihood-based algorithm whose empirical performance is evaluated through extensive simulation studies. Some theoretical and asymptotic results are discussed. Two case studies are presented using population-based cancer data from patients diagnosed with breast (female), colon and lung cancers in England. The results support the presence of non-linear and time-dependent effects as well as spatial variation.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

The proposed approach is available in the R package GJRM.

KEYWORDS

additive predictor, excess hazard, left truncation, link function, mixed censoring, net survival, penalized log-likelihood, regression splines, spatial effects, survival data

1 | INTRODUCTION

One of the aims of population cancer epidemiology consists of quantifying the survival due to cancer and to describe inequalities in cancer survival outcomes. This includes comparisons of cancer survival between different subgroups of the populations, such as those defined by different socio-economic or geographical factors. Cancer survival is typically used as a proxy for the overall effectiveness of the healthcare system in the treatment and management of cancer (Coleman, 2014), and it is increasingly used to formulate cancer control strategies (Department of Health, 2011). Data for cancer research are available from population-based cancer registries which collect a standard set of information for every cancer registration, covering patient demographics, tumour characteristics and type of treatment. Many efforts have been made in recent years to augment cancer registration data with relevant clinical information contained in other electronic health databases. Such enriched data create new opportunities for more complex cancer research questions to be investigated.

There are three main frameworks for analysing survival data. The first is the overall survival framework, where all-cause mortality is studied. This quantity is not of interest in cancer survival studies because it does not quantify the survival due to cancer. The second is the cause-specific framework, where information on the different causes of death is available, for example in the death certificates. This addresses the previous issue as it indeed accounts for the different causes of mortality in the population. Unfortunately, death certificates are unreliable in virtually any country in the world, at least at the population level. The third is the relative survival framework, which can be formulated in the absence of information on the cause of death. In this framework, the idea is to separate the hazard associated with other causes of death from that associated with cancer. This is done by assuming an additive decomposition of the individual hazard function, $h(\cdot)$, into two parts: the hazard associated with other causes of death, $h_O(\cdot)$, and the hazard associated with cancer, $h_E(\cdot)$ (Estève et al., 1990):

$$h(t|\mathbf{x}) = h_O(\text{age} + t) + h_E(t|\mathbf{x}), \tag{1}$$

where 'age' is the age at diagnosis of cancer and $\mathbf{x}$ represents the available patient characteristics. The hazard associated with other causes of death, $h_O(\text{age} + t)$, is typically replaced by the population hazard rate $h_P(\text{age} + t|\mathbf{w})$, which is obtained from life tables based on available characteristics denoted by the generic vector $\mathbf{w} \subset \mathbf{x}$ which can possibly include, in addition to age at death or censoring (age + t), gender and calendar year, socio-economic status, ethnicity or region of residence (Rachet et al., 2015). More specifically, $h_O(\text{age} + t)$ represents the true theoretical hazard function associated with other causes of death and as such it is unknown in practice. For this reason, it is approximated by $h_P(\text{age} + t|\mathbf{w})$, which can instead be extracted from national life

tables as mentioned above. As an aside, note that the true theoretical hazard function $h_O(\text{age} + t)$ may depend on $\mathbf{x}$, or on a subset of $\mathbf{x}$, or even on covariates that are not recorded.

The hazard associated with cancer, $h_E(t|\mathbf{x})$, is often referred to as the *excess hazard*. The excess hazard function is typically modelled using the available patient characteristics, denoted by $\mathbf{x}$ which can, for instance, incorporate continuous and categorical variables in our framework. Several approaches for estimating the excess hazard have been explored in the literature, such as non-parametric methods, which aim at estimating the cumulative excess hazard (Perme et al., 2012) and the net survival (Pavlič & Pohar-Perme, 2019; Pohar-Perme et al., 2009, 2016), parametric methods based on flexibly modelling the baseline excess hazard or cumulative hazard using splines (Charvat et al., 2016; Cramb et al., 2016; Fauvernier et al., 2019; Lambert & Royston, 2009; Quaresma et al., 2019) and modelling the baseline excess hazard function using flexible parametric distributions (Rubio et al., 2019). Most approaches assume a proportional hazards (PH) structure (with the option of adding time-dependent effects as originally proposed by Cox (1972), which is a convenient way of bypassing the proportionality assumed by the PH setting), with the exception of Rubio et al. (2019), who adopt a general hazard structure that contains the PH, accelerated hazards and the accelerated failure time (AFT) models as particular cases.

We propose a flexible parametric modelling framework. In this respect, it should be noted that non-parametric and parametric approaches are generally viewed as complementary by practitioners, rather than mutually exclusive. This view is strengthened by the fact that they are not directly comparable. The interpretation for non-parametric models is, in fact, different than for parametric models. This includes but is not limited to the fact that parametric approaches can account for covariates directly while non-parametric approaches cannot (Perme et al., 2012). Further, as mentioned above the available approaches do not allow one to model the excess hazard function, as they represent estimates of the cumulative hazard or net survival (Perme et al., 2012). Instead, parametric approaches allow one to estimate and plot the excess hazard function as this function is explicitly available (Rubio et al., 2019).

Finally, with regard to our choice of taking a parametric approach, we note that Cox has encouraged the broader use of parametric survival models for empirical modelling (Hjort, 1992; Reid, 1994). This is because they facilitate model estimation and comparison, easily allow for the calculation and visualization of, for instance, the estimated baseline hazard and survival functions, and allow one to calculate many quantities of interest and their related intervals (e.g. time-dependent hazard or odds ratios). Moreover, we overcome the generally restrictive nature of traditional parametric models by proposing a splines-based framework which allows for a great degree of modelling flexibility.

Based on the decomposition of the hazard function (1), the cumulative hazard function can be written as

$$H(t|\mathbf{x}) = \int_0^t h(r|\mathbf{x})dr = H_P(\text{age} + t|\mathbf{w}) - H_P(\text{age}|\mathbf{w}) + H_E(t|\mathbf{x}). \tag{2}$$

Consequently, the survival function can be factorized as follows:

$$S(t|\mathbf{x}) = \exp\{-H(t|\mathbf{x})\} = \exp\left\{-H_P(\text{age} + t|\mathbf{w}) + H_P(\text{age}|\mathbf{w})\right\} \exp\left\{-H_E(t|\mathbf{x})\right\}. \tag{3}$$

The survival function associated with the excess hazard, $S_N(t|\mathbf{x}) = \exp\{-H_E(t|\mathbf{x})\}$, is denoted as the (individual) *net survival*. The concept of net survival is usually favoured by international

agencies and programmes devoted to the study of cancer epidemiology, as well as policy-makers, as it is not affected by other causes of mortality, under the assumed model (1); we refer the reader to Rubio et al. (2019) and Rubio et al. (2021) for a discussion on these points.

Building on Marra and Radice (2020), we present a flexible methodology that is capable of handling simultaneously all types of censoring as well as left truncation, while accounting for the excess hazard. Often only right censoring and potentially left truncation is allowed (e.g. Fauvernier et al., 2019; Quaresma et al., 2019), thus accounting for any type of censoring broadens the applicability of our framework. Furthermore, a variety of covariate effects, including time-dependent effects, can be flexibly estimated via additive predictors with several types of smoothers. Our framework can also accommodate spatial effects in the definition of the additive predictor of an excess hazard model, a feature that is not available in other frameworks and software; this allows us to explore geographical disparities in cancer survival. The proposed model yields as special cases the widely used PH model, which allows for the usual Cox-like interpretation of the estimated effects, as well as the proportional odds (PO) model. The framework is based on modelling transformations of the survival function which we found to perform well in practice. An advantage of using this scale is that the post-estimation extraction of the (sub-)population net survival, a quantity often of interest to practitioners, is notably quicker when compared to approaches which model on the hazards scale; the need for numerical integration in the latter implies a higher computational time.

The resulting additive model is very flexible since the baseline hazard is modelled by means of monotonic P-splines. This is more efficient and parsimonious than using a non-parametric hazard, as in the Cox model, and it is more flexible than strong parametric assumptions such as those in AFT models. Parameter estimation is based on a penalized maximum likelihood approach that allows for stable and efficient computations where smoothness is guaranteed by means of a quadratic penalty. In order to allow for transparent and reproducible research as well as faster dissemination of scientific results in industry and academia, the proposed modelling framework is implemented in the `GJRM` `R` package (Marra & Radice, 2021). This implementation allows the applied end-user to obtain and visualize relevant quantities such as population net survival and excess hazard, and their confidence intervals, and easily perform model comparisons. Various examples of code usage can be found in the online Supplementary Material as well as on the public repository https://github.com/FJRubio67/LBANS/, where two publicly available data sets are analysed.

Sections 2 and 3 present the model formulation and the model's penalized log-likelihood. Section 4 discusses parameter estimation and inference as well as some theoretical results. Section 5 contains the results of the simulation study. Sections 6.1 and 6.2 present two case studies in the context of cancer epidemiology. Section 7 concludes the paper with a discussion and potential directions for future research. Finally, for the sake of space, several details are collected in the online Supplementary Material.

2 | FLEXIBLE EXCESS HAZARD MODEL

For individual $i = 1, \dots, n$, where n represents the sample size, let T_i denote the true event time and have a conditional net survival function denoted by $S_N(t_i|\mathbf{x}_i; \boldsymbol{\beta}) = \exp\{-H_E(t_i|\mathbf{x}_i; \boldsymbol{\beta})\} \in (0, 1)$, where $\mathbf{x}_i$ represents a generic vector of patient characteristics that has an associated regression

coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^w$, where w is the length of $\boldsymbol{\beta}$. A link-based additive net survival model can be written as

$$g\left\{S_N(t_i|\mathbf{x}_i;\boldsymbol{\beta})\right\} = \eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta})), \tag{4}$$

where $g : (0,1) \to \mathbb{R}$ is a monotone and twice continuously differentiable link function with bounded derivatives and hence invertible, $\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta})) \in \mathbb{R}$ is an additive predictor which includes a baseline function of time, or a stratified set of functions of time, and several types of covariate effects (see the next section), and $\mathbf{f}(\boldsymbol{\beta})$ is a vector function of $\boldsymbol{\beta}$ whose main role is to impose the monotonicity constraint, discussed in Section 3, needed when evaluating the baseline function of time contained in the additive predictor. Note that the choice for g determines the scale of the analysis (e.g. Liu et al., 2018).

Rearranging (4) yields $S_N(t_i|\mathbf{x}_i;\boldsymbol{\beta}) = G\left\{\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))\right\}$, where G is an inverse link function. The cumulative hazard and hazard functions, H and h, are defined as $H_E(t_i|\mathbf{x}_i;\boldsymbol{\beta}) = -\log\left[G\left\{\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))\right\}\right]$ and

$$h_E(t_i|\mathbf{x}_i;\boldsymbol{\beta}) = -\frac{G'\left\{\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))\right\}}{G\left\{\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))\right\}}\frac{\partial\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))}{\partial t_i}, \tag{5}$$

where $G'\left\{\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))\right\} = \partial G\left\{\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))\right\}/\partial\eta_i(t_i,\mathbf{x}_i;\mathbf{f}(\boldsymbol{\beta}))$. Table 1 displays the functions g, G and G' considered in this work.

2.1 | Additive predictor

For the sake of simplicity, the dependence on covariates and parameters has been dropped when discussing the construction of η_i. Since t_i can be treated as a regressor, we define an overall covariate vector $\mathbf{z}_i$ made up of $\mathbf{x}_i$ and t_i. An additive predictor allows for various types of covariate effects as well as their flexible functional form determination. An additive predictor is defined as

$$\eta_i = \beta_0 + \sum_{k=1}^{K} s_k(\mathbf{z}_{ki}), \quad i = 1, \ldots, n, \tag{6}$$

where $\beta_0 \in \mathbb{R}$ is an overall intercept, $\mathbf{z}_{ki}$ denotes the kth subvector of the complete vector $\mathbf{z}_i$ and the K functions $s_k(\mathbf{z}_{ki})$ denote effects which are chosen according to the type of covariate(s) considered. Each $s_k(\mathbf{z}_{ki})$ can be represented as a linear combination of J_k basis functions $b_{kj_k}(\mathbf{z}_{ki})$ and regression coefficients $f_{kj_k}(\beta_{kj_k}) \in \mathbb{R}$, that is (e.g. Wood, 2017)

TABLE 1 Functions implemented in `GJRM` (see Marra and Radice (2020) and references therein). Φ and ϕ are the cumulative distribution and density functions of a univariate standard normal distribution. The first two functions are typically known as log-log and -logit links, respectively

Model	Link $g(S)$	Inverse link $g^{-1}(\eta) = G(\eta)$	$G'(\eta)$
Prop. hazards (`"PH"`)	$\log\{-\log(S)\}$	$\exp\{-\exp(\eta)\}$	$-G(\eta)\exp(\eta)$
Prop. odds (`"PO"`)	$-\log\left(\frac{S}{1-S}\right)$	$\frac{\exp(-\eta)}{1+\exp(-\eta)}$	$-G^2(\eta)\exp(-\eta)$
probit (`"probit"`)	$-\Phi^{-1}(S)$	$\Phi(-\eta)$	$-\phi(-\eta)$

$$\sum_{j_k=1}^{J_k} f_{kj_k}(\beta_{kj_k}) b_{kj_k}(\mathbf{z}_{ki}). \tag{7}$$

The above formulation implies that the vector of evaluations $\{s_k(\mathbf{z}_{k1}), \ldots, s_k(\mathbf{z}_{kn})\}^\top$ can be written as $\mathbf{Z}_k \mathbf{f}_k(\boldsymbol{\beta}_k)$ with $\mathbf{f}_k(\boldsymbol{\beta}_k) = (f_{k1}(\beta_{k1}), \ldots, f_{kJ_k}(\beta_{kJ_k}))^\top$ and design matrix $\mathbf{Z}_k[i, j_k] = b_{kj_k}(\mathbf{z}_{ki})$. This allows the predictor in Equation (6) to be written as

$$\boldsymbol{\eta} = \beta_0 \mathbf{1}_n + \mathbf{Z}_1 \mathbf{f}_1(\boldsymbol{\beta}_1) + \ldots + \mathbf{Z}_K \mathbf{f}_K(\boldsymbol{\beta}_K), \tag{8}$$

where $\mathbf{1}_n$ is an n-dimensional vector made up of ones. Equation (8) can also be written in a more compact way as $\boldsymbol{\eta} = \mathbf{Z}\mathbf{f}(\boldsymbol{\beta})$, where $\mathbf{Z} = (\mathbf{1}_n, \mathbf{Z}_1, \ldots, \mathbf{Z}_K)$ and $\mathbf{f}(\boldsymbol{\beta}) = (\beta_0, \mathbf{f}(\boldsymbol{\beta}_1)^\top, \ldots, \mathbf{f}(\boldsymbol{\beta}_K)^\top)^\top$. Additional observations on the additive predictor described here can be found in online Supplementary Material C.1.

Each $\boldsymbol{\beta}_k$ has an associated quadratic penalty $\lambda_k \boldsymbol{\beta}_k^\top \mathbf{D}_k \boldsymbol{\beta}_k$, used in fitting, whose role is to enforce specific properties on the kth function, such as smoothness. Note that matrix $\mathbf{D}_k$ only depends on the choice of the basis functions. The smoothing parameter $\lambda_k \in [0, \infty)$ controls the trade-off between fit and smoothness, and hence determines the shape of the estimated smooth function. The overall penalty can be defined as $\boldsymbol{\beta}^\top \mathbf{S} \boldsymbol{\beta}$, where $\mathbf{S} = \text{diag}(0, \lambda_1 \mathbf{D}_1, \ldots, \lambda_K \mathbf{D}_K)$. Note that smooth functions are subject to centring (identifiability) constraints which can be imposed as described in Wood (2017). Depending on the types of covariate effects one wishes to model, several definitions of basis functions and penalty terms are possible. Examples include thin plate, cubic and P- regression splines, tensor products, Markov random fields (MRFs), random effects, Gaussian process smooths (see Wood, 2017, for all the options available). More details can be found in the case studies reported in Section 6.

Finally, observe that in Equation (5) quantity $\partial \eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))/\partial t_i$ is required. Re-writing $\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))$ as $\mathbf{Z}_i(t_i, \mathbf{x}_i)^\top \mathbf{f}(\boldsymbol{\beta})$, the derivative of interest can be obtained as $\lim_{\varepsilon \to 0} \left\{ \frac{\mathbf{Z}_i(t_i+\varepsilon, \mathbf{x}_i) - \mathbf{Z}_i(t_i-\varepsilon, \mathbf{x}_i)}{2\varepsilon} \right\}^\top \mathbf{f}(\boldsymbol{\beta}) = \mathbf{Z}_i'^\top \mathbf{f}(\boldsymbol{\beta})$, where, depending on the type of spline basis employed, $\mathbf{Z}_i'$ can be calculated either by a finite-difference method or analytically.

3 | PENALIZED LOG-LIKELIHOOD

The unifying framework proposed in this paper supports excess hazard modelling, all types of censoring and left truncation in addition to the flexible additive predictor introduced in Section 2.1. As the case studies presented in Section 6 involve excess hazard modelling on right-censored data, which is the most common scenario in cancer research, here we will define the setting and the log-likelihood only for this case. A detailed discussion of the full log-likelihood for the general case can be found in online Supplementary Material A, while its derivation is reported in online Supplementary Material B.

When the ith true event time T_i is known exactly, the individual is said to be uncensored. In some cases, however, T_i may only be known to be larger than a certain time R_i, in which case the individual is said to be right-censored and R_i is the random right-censoring time. The censoring type of the ith observation can be summarized through the use of the indicator functions δ_{Ri} and δ_{Ui}, where $\delta_{Ri} = 1$ if the observation is right censored and 0 otherwise while $\delta_{Ui} = 1$ if it is uncensored and 0 otherwise.

Let us assume that a random *i.i.d.* sample $\{(r_i, \delta_{Ui}, \delta_{Ri}, \mathbf{x}_i)\}_{i=1}^n$ is available, where r_i is either the time of death or the observed right-censoring time, and that censoring is independent and non-informative conditional on $\mathbf{x}_i$. Let us also write $S_N(t_i|\mathbf{x}_i) = S_N\{\eta_i(t_i)\}$ in order to make the dependence of the net survival on η explicit. The log-likelihood function associated with the additive excess hazard model (1)–(3) can be written as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \delta_{Ui} \log\left[h_P(\text{age}_i + r_i|\mathbf{w}_i) S_N\{\eta_i(r_i)\} - \frac{\partial S_N\{\eta_i(r_i)\}}{\partial \eta_i(r_i)} \frac{\partial \eta_i(r_i)}{\partial r_i}\right] + \sum_{i=1}^n \delta_{Ri} \log[S_N\{\eta_i(r_i)\}] + C_i, \tag{9}$$

where r_i is the exact event time when $\delta_{Ui} = 1$ and where C_i is a constant with respect to the model's parameters whose expression can be found in online Supplementary Material B.

The proposed model allows for a high degree of flexibility, which is why penalized estimation of $\boldsymbol{\beta}$ is advisable. In order to prevent over-fitting, we maximize the penalized log-likelihood

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^\top \mathbf{S}\boldsymbol{\beta}. \tag{10}$$

To ensure that the estimated survival function is monotonically decreasing or equivalently that the hazard function is positive, the time effects are modelled using the monotonic P-spline approach. Let $s(t_i) = \sum_{j=1}^J f_j(\beta_j) b_j(t_i)$, where the b_j are B-spline basis functions of at least second order built over the interval $[a, b]$, based on equally spaced knots, and the $f_j(\beta_j)$ are spline coefficients. A sufficient condition for $s'(t_i) \geq 0$ over $[a, b]$ is that $f_j(\beta_j) \geq f_{j-1}(\beta_{j-1}), \forall j$ (e.g. Leitenstorfer & Tutz, 2007). Such condition can be imposed by defining the vector function of $\boldsymbol{\beta}$ as $\mathbf{f}(\boldsymbol{\beta}) = \boldsymbol{\Sigma}\{\beta_1, \exp(\beta_2), \ldots, \exp(\beta_J)\}^\top$, where $\boldsymbol{\Sigma}[\iota_1, \iota_2] = 0$ if $\iota_1 < \iota_2$ and $\boldsymbol{\Sigma}[\iota_1, \iota_2] = 1$ if $\iota_1 \geq \iota_2$, with ι_1 and ι_2 denoting the row and column entries of $\boldsymbol{\Sigma}$, and $\boldsymbol{\beta}^\top = (\beta_1, \beta_2, \ldots, \beta_J)$ is the parameter vector to estimate. Note that in practice $\boldsymbol{\Sigma}$ is absorbed into the design matrix containing the B-spline basis functions $\mathbf{Z}$. When setting up the penalty term, we penalize the squared differences between adjacent β_j, starting from β_2, using $\mathbf{D} = \mathbf{D}^{*\top}\mathbf{D}^*$ where $\mathbf{D}^*$ is a $(J-2) \times J$ matrix made up of zeros except that $\mathbf{D}^*[\iota, \iota+1] = -\mathbf{D}^*[\iota, \iota+2] = 1$ for $\iota = 1, \ldots, J-2$ (Pya & Wood, 2015).

4 | PARAMETER ESTIMATION AND INFERENCE

The estimation approach employed in this article is based on analytical derivative information which helps enhance numerical stability and speed. It is worth noting that, given the structure of (9), deriving such quantities has been a tedious task. Furthermore, the above mentioned re-parametrization implies a non-linear dependence of $\mathbf{f}(\boldsymbol{\beta})$ from $\boldsymbol{\beta}$ which additionally complicates the structure of the derivatives, in particular those of the additive predictor $\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))$ with respect to parameter vector $\boldsymbol{\beta}$. These appear repeatedly in the score and in the Hessian and are given by $\partial\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))/\partial\boldsymbol{\beta} = \mathbf{Z}_i \circ \mathbf{E}$, $\partial^2\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))/\partial\boldsymbol{\beta}\partial t_i = \mathbf{Z}_i' \circ \mathbf{E}$ and $\partial^2\eta_i(t_i, \mathbf{x}_i; \mathbf{f}(\boldsymbol{\beta}))/\partial\boldsymbol{\beta}^2 = \text{diag}(\mathbf{Z}_i) \circ \bar{\bar{\mathbf{E}}}$, where $\mathbf{Z}_i$ is the transformed covariate vector corresponding to the ith observation, $\circ$ is the Hadamard product, diag($\mathbf{v}$) is a diagonal matrix with $\mathbf{v}$ its diagonal, $\mathbf{E}$ is a $\sum_{k=1}^K J_k \times 1$ vector such that its $k_{j_k}^{th}$ element is $\mathbf{E}[k_{j_k}] = 1$ if $f_{k_{j_k}}(\beta_{k_{j_k}}) = \beta_{k_{j_k}}$ and $\exp(\beta_{k_{j_k}})$ otherwise, and $\bar{\bar{\mathbf{E}}}$ is a

$\sum_{k=1}^{K} J_k \times \sum_{k=1}^{K} J_k$ diagonal matrix such that its $k_{j_k}^{th}$ diagonal element is $\bar{\bar{\mathbf{E}}}[k_{j_k}, k_{j_k}] = 0$ if $f_{k_{j_k}}(\beta_{k_{j_k}}) = \beta_{k_{j_k}}$ and $\exp(\beta_{k_{j_k}})$ otherwise.

The analytical expression of the gradient and Hessian matrix are presented in online Supplementary Material B. Although these derivatives involve lengthy calculations as well as careful algorithmic implementation, the computational and inferential benefits of avoiding numerical approximations justify the effort. The algorithm employed for estimating the regression parameters and smoothing coefficient vector is summarized in online Supplementary Material C.2. Briefly, it combines a carefully structured trust region algorithm which uses the analytical expressions of the gradient and Hessian of the log-likelihood and properly chosen starting values with a general automatic multiple smoothing parameter selection algorithm based on an approximate AIC measure.

In practice, this results in an estimation algorithm which is general, modular, efficient and stable, working well even for problems which are non-concave and/or exhibit close to flat regions. We found this both through usage on real-world data as well as through the extensive simulation study conducted and reported in detail in online Supplementary Material F. As expected, like any method, in the latter we found that model fitting failed to converge at times (i.e. did not achieve close to zero gradient and/or positive definite Hessian), however this occurred only for a small percentage of simulation replicates. This is in line with what we found with `survPen`, our main competitor. Further details on this can also be found in online Supplementary Material F.

To obtain confidence intervals, we follow Wood et al. (2016) and employ the Bayesian large sample approximation $\boldsymbol{\beta} \dot{\sim} \mathcal{N}(\widehat{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$, where $\mathbf{V}_{\boldsymbol{\beta}} = -\boldsymbol{H}_p(\widehat{\boldsymbol{\beta}})^{-1}$; using $\mathbf{V}_{\boldsymbol{\beta}}$ gives close to across-the-function frequentist coverage probabilities because it accounts for both sampling variability and smoothing bias, a feature that is particularly relevant at finite sample sizes. Note that applying the Bayesian approach to the modelling framework discussed in this paper follows the notion that penalization in estimation implicitly assumes that wiggly models are less likely than smoother ones, which translates into the following prior specification for $\boldsymbol{\beta}$, $f_{\boldsymbol{\beta}} \propto \exp\{-\boldsymbol{\beta}^\top \mathbf{S}\boldsymbol{\beta}/2\}$.

Since the evaluation of the additive predictor in Equation (8) and the quantities that rely on it depend on $\mathbf{f}(\boldsymbol{\beta})$, it makes sense to obtain its distribution as well. Following Pya and Wood (2015), we first consider the Taylor series expansion of $\mathbf{f}(\boldsymbol{\beta})$ around $\mathbf{f}(\tilde{\boldsymbol{\beta}})$, that is $\mathbf{f}(\boldsymbol{\beta}) - \mathbf{f}(\tilde{\boldsymbol{\beta}}) \approx \mathrm{diag}(\mathbf{E})\left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)$. This shows that $\mathbf{f}(\boldsymbol{\beta}) - \mathbf{f}(\tilde{\boldsymbol{\beta}})$ is approximately a linear function of $\boldsymbol{\beta}$. We then recall that linear functions of normally distributed random variables follow normal distributions. This implies that $\mathbf{f}(\boldsymbol{\beta}) \dot{\sim} \mathcal{N}(\mathbf{f}(\tilde{\boldsymbol{\beta}}), \mathbf{V}_{\mathbf{f}(\boldsymbol{\beta})})$ where $\mathbf{V}_{\mathbf{f}(\boldsymbol{\beta})} = \mathrm{diag}(\mathbf{E})\mathbf{V}_{\boldsymbol{\beta}}\mathrm{diag}(\mathbf{E})$. p-values for the smooth components in the model are derived by adapting the result discussed in Wood (2017) and using $\mathbf{V}_{\mathbf{f}(\boldsymbol{\beta})}$ as covariance matrix.

Intervals for linear functions of the model's coefficients, for example smooth components, can then be obtained using the result just shown for $\mathbf{f}(\boldsymbol{\beta})$. For non-linear functions of the model's coefficients, for example hazard functions, instead, the intervals can be conveniently obtained by posterior simulations, hence avoiding computationally expensive parametric bootstrap or frequentist approximations, for instance.

The approximation found for $\mathbf{f}(\boldsymbol{\beta})$ also facilitates the construction of confidence intervals for the net survival curve (either associated with an individual or a subpopulation). We define the (marginal) net survival function associated with a subpopulation $\mathbf{x}_{pop} = \{\mathbf{x}_1, \ldots \mathbf{x}_k\}$ as

$$\overline{S}_N(t) = \frac{1}{k} \sum_{\mathbf{x}_i \in \mathbf{x}_{pop}} S_N(t|\mathbf{x}_i),$$

where it is assumed that k is the number of individuals belonging to the subpopulation of interest. For instance, $\mathbf{x}_{pop}$ could be the entire population or a subgroup of interest, such as a specific age group. Keeping in mind that we are interested in finding the interval for an average over multiple net survival curves, we will have to sample from the posterior distribution of this average. Finally, online Supplementary Material D describes the use of the `R` package `GJRM`.

The main asymptotic results related to the proposed estimator are presented below and are based on classical assumptions from the GAM and relative survival literature and refer to model regularity conditions.

Theorem 1 *If Assumptions A1–A8 hold (see online Supplementary Material E) then*

1. $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ *and* $||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| = O_p(n^{-\frac{1}{2}})$,
2. $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{i}^{-1}(\boldsymbol{\beta}_0))$,

where $\boldsymbol{\beta}_0$ *is the true parameters vector.*

5 | SIMULATION STUDY

We consider 20 scenarios, resulting from the combination of four sample sizes, $n = 200$, 500, 1000 and 5000, and five data generating processes (DGPs) of increasing difficulty, to extensively test our method's ability to capture the true generating mechanism. We will compare the performance of our method, implemented in the `R` package `GJRM`, with a state-of-the-art model in the (relative) survival setting, that is Fauvernier et al. (2019) and its implementation in the `R` package `survPen`. Although other penalized relative survival model implementations exist, we consider `survPen` to be an adequate benchmark as it was in turn extensively tested against competing frameworks in the reference paper and was generally found to be superior. We will then have four fitted models: `GJRM` with each of the three allowed link functions, that is PH, PO and probit, and `survPen`.

As we are in a relative survival setting, we will simulate the population hazard and the excess hazard separately for each individual. The former is simulated from a piece-wise exponential distribution based on life tables from the general English population. The latter using increasingly complex functional forms with parameters set to result in approximately 40% censoring. This level was chosen to reflect the 44.8% censoring found in the case studies on which the simulated ones are based. For each scenario, we simulate 1000 data sets, which include also age at diagnosis and level of deprivation defined on a discrete scale between 1 (least deprived) and 5 (most deprived).

Our method performs consistently well throughout the scenarios and over a range of metrics, with greater uncertainty generally found at the smaller sample size, as expected. In the following section, we present the results for one of the most challenging DGPs. For more details and the full set of results, we refer the reader to online Supplementary Material F.

5.1 | General hazards model with non-linear effect of age

We consider a general hazards model, as defined in Rubio et al. (2019). For the jth observation this is given by

$$h_E^{GH}(t;\mathbf{x}_j) = h_0\left(t\exp[\alpha \cdot f(\text{agec}_j)]\right)\exp(\mathbf{x}_j^\top \boldsymbol{\beta}) \quad \text{with} \quad h_0(r) = \frac{\phi\left(\frac{\ln r-\mu}{\sigma}\right)}{\Phi\left(-\frac{\ln r-\mu}{\sigma}\right)\cdot r\sigma},$$

where the baseline hazard $h_0(\cdot)$ is modelled using a log-normal distribution with parameters (μ, σ) and where $\phi(\cdot)$ and $\Phi(\cdot)$ represent, respectively, the density function and CDF of a standard normal. Furthermore, $\mathbf{x}_j = [f(\text{agec}_j), \text{dep}_j]^\top$ where $f(\text{agec}_j) = 0.75 \sinh[0.5\text{arcsinh}[3\text{agec}_j]]$ is a smooth function of the standardized age at diagnosis agec_j, chosen to ensure that the associated effect was not too small, and dep_j is the level of deprivation. The associated (time-fixed) effect is denoted by $\boldsymbol{\beta}$. A time-dependent effect is also assumed for $f(\text{agec}_j)$ and is denoted by α. The values used for the parameters $(\mu, \sigma, \alpha, \boldsymbol{\beta})$ are reported in Table 8 of online Supplementary Material F.

In terms of model selection, using the AIC, the `GJRM` PH model was found to be the best among the `GJRM` models which were in turn found to be preferred to `survPen`. This holds with the exception of the case $n = 200$, for which `GJRM` probit is the best model in terms of AIC, both when compared with the other `GJRM` models as well as with `survPen`; see table 10 of the online Supplementary Material F. In the following, only the best `GJRM` model will be compared with the `survPen` model, to avoid cluttering the plots.

In Figure 1, we report the boxplot of the root mean square error (RMSE) for the excess hazard for each of the two approaches and for each sample size. The boxplot of the bias is very similar so it is omitted here due to space constraints but it is reported in Figure 22 of online Supplementary Material F. We find that the RMSE decreases as the sample size increases and that it is overall smaller but more variable for `GJRM` PH than for `survPen` at $n = 1000$ and $n = 5000$. At $n = 200$ and $n = 500$, `survPen` outperforms `GJRM` probit and `GJRM` PH, respectively; from Figure 2, we see that the `survPen` estimated excess hazard curve is not very close to the true curve although it is overall better than the one produced by `GJRM`. In general, `GJRM` mostly leads to curves which trace the true excess hazard more closely but, when they do not, specifically in the middle and final times, they contribute to higher overall values of RMSE. We find this behaviour in the average estimated excess hazard plots reported in Figure 2, where `GJRM` captures the first portion of the true excess hazard relatively well, even at the lowest sample sizes, but departs from it in the final times. `survPen`, instead, struggles to capture the initial steeply increasing portion of the true excess hazard while it is closer to the true curve in the final times. At the highest sample size, `GJRM` improves greatly its fit also in the final times, becoming almost indistinguishable from the

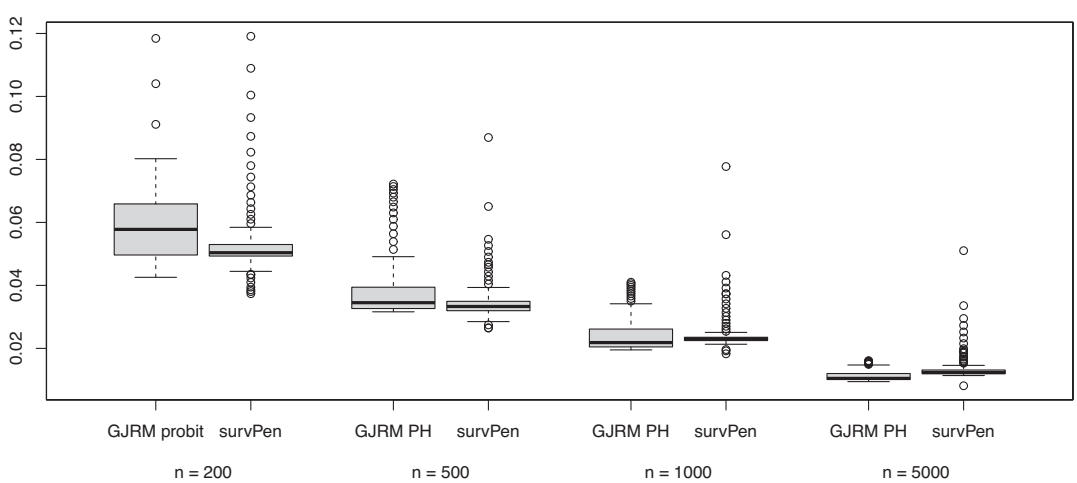
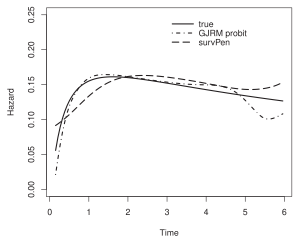
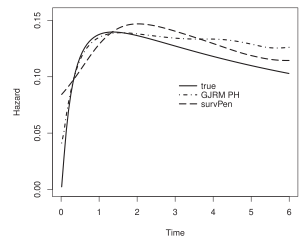


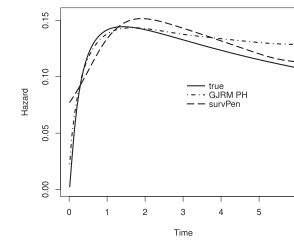
FIGURE 1 Boxplots of root mean square error of excess hazard function for each model specification under the fourth data generating processes



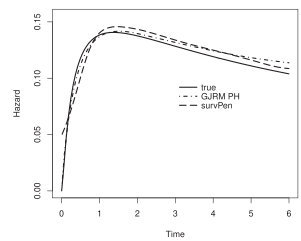
(a) Sample size $n = 200$.



(b) Sample size $n = 500$.



(c) Sample size $n = 1000$.



(d) Sample size $n = 5000$.

FIGURE 2 Average estimated excess hazard for `GJRM` and `survPen` versus the true excess hazard function under the fourth data generating processes (the mean was taken across the excess hazard estimated on each simulated data set)

true curve across the entire time range. `survPen` also greatly improves its fit to the first portion of the true curve but does not match `GJRM`.

We omit the analysis of the population net survival as both methods perform very well, with almost perfectly overlapping estimated average curves when compared with the true survival. Consequently, the bias and RMSE are markedly smaller than for the estimated excess hazard with the two methods overall comparable in magnitude and variability. These details can nonetheless be found in online Supplementary Material F.

In conclusion, the simulation study shows that `GJRM` and `survPen` are comparable and both able to adequately capture even complex DGPs. In particular, `survPen` seems to perform better when n is smaller and, while capturing satisfactorily the overall shape of the true excess hazard, it appears slightly shifted and flatter. `GJRM` performs less well at the lowest sample size but the fit improves dramatically as the sample size increases. As all of the data sets considered in the applications are characterized by a number of patients which is well above the largest sample size considered here, empirical performance is not likely to be a matter of concern. Finally, both approaches are characterized by similar model specifications and run times for model fitting, however `survPen` is slower when calculating (sub-)population net survival estimates. For more details on these aspects and for the full set of results, we refer the reader to online Supplementary Material F.

6 | CASE STUDIES

Cancer research strives to provide an accurate picture of the evolving cancer burden, as well as documenting existing inequalities, using a variety of key indicators, including cancer survival. In

this section, we present two case studies which aim at investigating inequalities in net survival for patients diagnosed in England. The first case study aims at investigating socio-economical inequalities in net survival for the top three incident cancer types (breast, colon and lung), but which have differential levels of survival as confirmed by previous research (Quaresma et al., 2015), and the second case study aims at studying geographical disparities in net survival for colon cancer patients in England.

For these case studies, individual cancer records were obtained from the National Cancer Registry at the Office for National Statistics (ONS) on all adult patients (aged 15–99 years) diagnosed with a first, primary, invasive malignancy of the breast (women), colon and lung during 2010 in England. All cancer records were followed-up by the National Health Service Central Register, who routinely update these records with information about each patient's vital status. These cases were followed up until the 31 December 2015. Survival times were measured from the date of diagnosis until the date of death or last time of follow-up. The individual patient and tumour-specific variables available for this analysis were: full date of diagnosis, last follow-up and death times, vital-status indicator (which takes value 1 if the patient died of the cancer of interest and 0 otherwise), age at diagnosis (recorded as a continuous variable) and deprivation category (1—least deprived to 5—most deprived) defined according to the quintiles of the distribution of the Income Domain scores of the 2011 England Indices of Multiple Deprivation, NHS England Regions and Local Offices of residence (14 geographical regions), and tumour stage at diagnosis (I–IV). Background mortality rates were obtained for each cancer patient from population life tables for England defined for each calendar year in 2010–2015, and stratified by single year, age, sex, deprivation category and Government Office Region of residence.

Additional examples of data analyses, including of spatial effects on cancer survival, can be found on the public repository https://github.com/FJRubio67/LBANS/, where we consider the Simulacrum data set and the `LeukSurv` data set from the `R` package `spBayesSurv`.

6.1 | Socio-demographic inequalities in breast, colon and lung cancer survival in England

In this case study, we compare the net survival curves for the most deprived and least deprived groups of the population for the three major cancer types. The breast cancer in women data set contains $n = 38{,}636$ complete cases, with median age 62.8 years, from which $n_o = 9169$ patients died within the follow-up period (76.2% censoring). The colon cancer in men data set contains $n = 11{,}106$ complete cases, with median age 72.7 years, from which $n_o = 6126$ patients died within the follow-up period (44.8% censoring). The colon cancer in women data set contains $n = 9{,}999$ complete cases, with median age 74.8 years, from which $n_o = 5520$ patients died within the follow-up period (44.8% censoring). The lung cancer in men data set contains $n = 18{,}609$ complete cases, with median age 72.8 years, from which $n_o = 17{,}286$ patients died within the follow-up period (7.1% censoring). The lung cancer in women data set contains $n = 14{,}920$ complete cases, with median age 73.2 years, from which $n_o = 13{,}418$ patients died within the follow-up period (10.1% censoring). We chose to analyse these data sets as cancers they involve are the three most commonly diagnosed types in England (Office for National Statistics, Newport, UK, 2011) with each having differential levels of survival. Previous research investigating trends in cancer survival in England since the 1970s has, in fact, identified three broad groups of cancers based on their levels of survival: those with good prognosis, including breast cancer, those with moderate

survival levels, including colon cancer, and those with very poor prognosis, including lung cancer, for which little improvement has occurred in the past 40 years up to 2010 (Quaresma et al., 2015).

Table 2 shows the net survival estimates and the corresponding 95% confidence intervals, at 1,3 and 5 years after diagnosis, using `GJRM`. Models equivalent to those specified for `GJRM` were fitted using `survPen` as well; the AIC favoured the `GJRM` models in all cases. The net survival estimates obtained using `survPen` were very close to those obtained with the best `GJRM` model; they can be found in table 13 of online Supplementary Material G.

Note, in particular, that the results reported in Table 2 correspond to the best model selected using the AIC among nine models obtained by combining the three different allowed links (PH, PO and probit) with three different definitions of the additive predictor. The first of these specifications includes a linear effect of age at diagnosis and takes the form $\eta_i = \beta_0 + \text{dep}_i^\top \boldsymbol{\beta}_1 + \text{agec}_i \beta_2 + s_1(\log(t_i))$, the second includes a non-linear effect of age at diagnosis and takes the form $\eta_i = \beta_0 + \text{dep}_i^\top \boldsymbol{\beta}_1 + s_1(\log(t_i)) + s_2(\text{agec}_i)$ while the third one includes a non-linear and time-dependent effect of age at diagnosis and takes the form $\eta_i = \beta_0 + \text{dep}_i^\top \boldsymbol{\beta}_1 + s_1(\log(t_i)) + s_2(\text{agec}_i) + s_3(\log(t_i), \text{agec}_i)$. Here dep_i represents the level of deprivation, defined on a discrete

TABLE 2 Net survival at 1, 3 and 5 years after diagnosis ('yrs') with 95% confidence interval between brackets for all adult (aged 15–99 years) women diagnosed with breast cancer and men and women diagnosed with colon and lung cancer during 2010 in England: population net survival ('pop'), net survival for the least deprived patients ('dep 1') and net survival for the most deprived patients ('dep 5'). The estimates were obtained using the `R` package `GJRM`

(yrs)	pop	dep 1	dep 5
Breast cancer			
1	0.96 (0.96, 0.96)	0.97 (0.97, 0.97)	0.95 (0.94, 0.95)
3	0.90 (0.89, 0.90)	0.92 (0.91, 0.92)	0.87 (0.87, 0.88)
5	0.85 (0.85, 0.86)	0.88 (0.87, 0.89)	0.82 (0.81, 0.83)
Colon cancer (men)			
1	0.74 (0.74, 0.75)	0.77 (0.76, 0.78)	0.71 (0.69, 0.72)
3	0.62 (0.61, 0.62)	0.65 (0.63, 0.66)	0.57 (0.55, 0.59)
5	0.57 (0.56, 0.58)	0.60 (0.58, 0.62)	0.52 (0.50, 0.54)
Colon cancer (women)			
1	0.72 (0.71, 0.73)	0.76 (0.75, 0.78)	0.68 (0.66, 0.69)
3	0.60 (0.58, 0.60)	0.64 (0.63, 0.66)	0.54 (0.52, 0.56)
5	0.55 (0.54, 0.56)	0.60 (0.59, 0.62)	0.50 (0.47, 0.52)
Lung cancer (men)			
1	0.30 (0.29, 0.30)	0.31 (0.30, 0.32)	0.29 (0.28, 0.30)
3	0.13 (0.12, 0.13)	0.13 (0.13, 0.14)	0.12 (0.12, 0.13)
5	0.09 (0.09, 0.09)	0.10 (0.09, 0.10)	0.09 (0.08, 0.09)
Lung cancer (women)			
1	0.34 (0.34, 0.35)	0.36 (0.35, 0.37)	0.33 (0.32, 0.34)
3	0.16 (0.15, 0.16)	0.17 (0.16, 0.18)	0.15 (0.14, 0.15)
5	0.12 (0.12, 0.12)	0.13 (0.12, 0.14)	0.11 (0.11, 0.12)

scale from 1 (least deprived) to 5 (most deprived), $s_1(\cdot)$ a monotonic P-spline taken over the logarithm of time chosen to ensure the monotonicity of the survival function as explained in Section 3, $s_2(\cdot)$ a cubic regression spline taken over the standardized age at diagnosis of cancer agec_i and $s_3(\cdot)$ a pure tensor product interaction between standardized age at diagnosis and time, whose marginals are also cubic regression splines. This is how time-dependent effects are included in the models specified in the applications and in the simulation study. Note that the term 'pure' refers to the fact that sum-to-zero constraints remove the unit function from the span of the marginals, with the result that the tensor product basis will not include the main effects. These, in fact, would result from the product of a marginal basis with the unit functions in the other marginal bases. In other terms, this specification enables us to model the main effects and the interaction term separately, thus leading to more flexibility as the main effects are allowed to have different complexity from their associated effects in the interaction term (Wood, 2017). With regard to the penalty associated with the non-linear term $s_2(\text{agec}_i)$, this takes the form of the quadratic penalty defined in Section 2.1 with $\mathbf{D}_k$ given by the integrated square second derivative of the basis functions, that is $\int \mathbf{d}_k(z_k)\mathbf{d}_k(z_k)^\top dz_k$ with the j_k^{th} element of $\mathbf{d}_k(z_k)$ defined as $\partial^2 b_{kj_k}(z_k)/\partial z_k^2$. The penalty associated with the non-linear pure interaction term $s_3(\log(t_i), \text{agec}_i)$ is, instead, more complex as it entails combining two penalties, each corresponding to one of the arguments of the smooth function. These are summed after being weighted by smoothing parameters, which thus serve the purpose of controlling the trade-off between the smoothness in each of the two directions; for more details on this we refer the reader to chapter 5 of Wood (2017). The best model according to the AIC is the one obtained by combining the probit link with the last specification. This is the case for all five data sets.

Figure 3 presents the net survival and the population excess hazard curves for data on breast, colon and lung cancer for female patients with deprivation categories 1 and 5. For completeness, we present the output for the best model as well as the smooths of age and $\log(t_i)$ for the colon cancer in men data set in online Supplementary Material G. Here, further details on how the non-linear and time-dependent effects can be specified in R can also be found.

Very high levels of survival were observed for women diagnosed with breast cancer in 2010 (above 80% at 5 years after diagnosis), moderate levels of survival for both men and women diagnosed with colon cancer (above 50% at 5 years after diagnosis) and very low survival for patients diagnosed with lung cancer in both genders. For all cancers, net survival was always lower for the most deprived group of patients at all times after diagnosis (Table 2). From the three cancers,

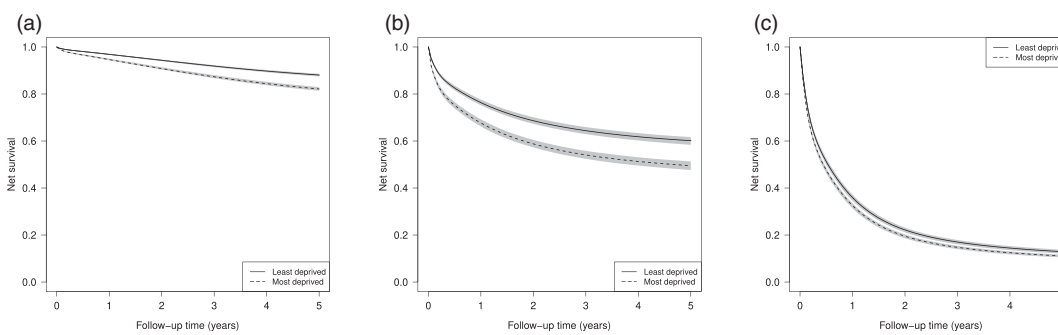


FIGURE 3 Net survival for all adult women (aged 15–99 years) diagnosed during 2010 in England (least deprived vs. most deprived): (a) breast cancer; (b) colon cancer; (c) lung cancer

the largest differences between the most deprived and the most affluent groups were observed for colon cancer patients.

6.2 | Modelling spatial effects on colon cancer survival in England

For this second case study, we highlight how to incorporate spatial effects in our methodology to analyse geographical inequalities in net survival. The data set contains $n = 9379$ complete cases, with median age 72.08 years, from which $n_o = 4859$ patients died within the follow-up period (48.2% censoring). There were 2092 patients with deprivation level 1 (least deprived), 2126 with deprivation level 2, 1946 with deprivation level 3, 1748 with deprivation 4 and 1467 with deprivation level 5 (most deprived). Among all patients, 1316 were diagnosed with stage 1 tumour, 2880 with stage 2, 2779 with stage 3 and 2404 with stage 4. We use as the geographical unit of analysis the NHS England Regions and Local Offices of residence.

We fitted the same nine models described in Section 6.1 with the only difference being that the additive predictors now include a spatial variation term as well. The best model in terms of AIC is that obtained by combining a PH link function with the most complex additive predictor specification, that is $\eta_i = \beta_0 + \text{dep}_i^\top \boldsymbol{\beta}_1 + s_1(\log(t_i)) + s_2(\text{agec}_i) + s_3(\log(t_i), \text{agec}_i) + s_{\text{spatial}}(\text{region}_i)$, where $s_{\text{spatial}}(\cdot)$ models the English National Health Service (NHS) regions and local offices of residence for individual i, indicated by region_i, using an MRF approach. In practice, considering R distinct regions, (7) takes the form $s_{\text{spatial}}(\text{region}_i) = \boldsymbol{\beta}_k^\top \boldsymbol{\delta}_i^{(reg)}$ where $\boldsymbol{\beta}_k = [\beta_{k1}, \dots, \beta_{kR}]^\top$ is the vector of effects associated with each region and $\boldsymbol{\delta}_i^{(reg)} = [\delta_{i1}^{(reg)}, \dots, \delta_{iR}^{(reg)}]^\top$ is such that $\delta_{ir}^{(reg)} = 1$ if individual i belongs to region r and 0 otherwise, for every $i = 1, \dots, n$ and $r = 1, \dots, R$. To ensure that neighbouring regions have similar effects β_{kr}, we penalize the sum of squared differences between β_{kr} values for all pairs of neighbouring regions. In other terms, we impose the penalty

$$\text{Pen}(\boldsymbol{\beta}_k) = \sum_{r=1}^{m} \sum_{\substack{q \in \text{nei}(r) \\ q > r}} (\beta_{kr} - \beta_{kq})^2,$$

where taking only the terms $q > r$ in the inmost summation ensures that the squared difference between a given pair is taken only once. Furthermore, $\text{nei}(r)$ represents the set of neighbours for region r. This can be re-written in terms of the quadratic penalty introduced above by defining an $R \times R$ matrix $\mathbf{D}_k$ with diagonal elements n_r given by the number of neighbours for region r and off-diagonal elements $\mathbf{D}_k[q, r] = -1$ if $q \in \text{nei}(r)$ and 0 otherwise, for $q, r = 1, \dots, R$. Note that the penalty can also be viewed as being induced by an improper Gaussian prior $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \tau \mathbf{D}_k^-)$, where τ is some precision parameter which replaces the smoothing parameter λ_k from the penalized likelihood framework through the equality $\tau = \lambda_k^{-1}$. The $\boldsymbol{\gamma}$ and the neighbourhood structure can then be viewed as an intrinsic Gaussian MRF with precision matrix $\mathbf{D}_k$ (Rue & Held, 2005).

The setup of the spatial effects in `R` is straightforward. The region boundaries, in fact, are openly accessible on the Office for National Statistics website. The regions can then be setup using the `GJRM` function `polys.setup()` and used in the model specification as the argument of the MRF smooth. For further details on how the models have been specified using the `R` package `GJRM` we refer the reader to online Supplementary Material G.

We report net survival at 1 and 5 years after diagnosis in Figure 4: Figures 4a and b present the results for the least deprived patients, and Figures 4c and d for most deprived patients. These plots have been obtained using the `GJRM` function `polys.map()`. In line with the

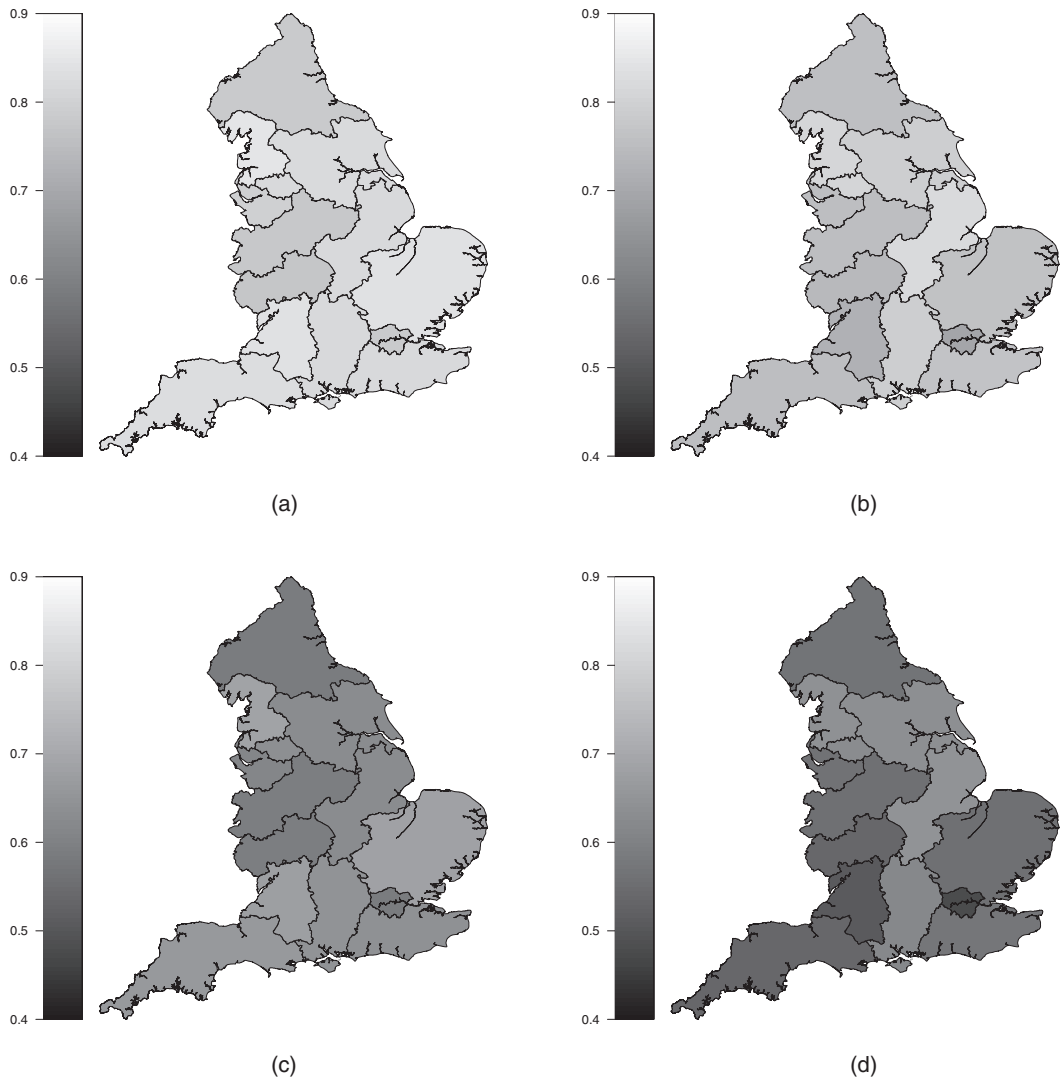


FIGURE 4 Maps of net survival for all adult male patients (aged 15–99 years) diagnosed with colon cancer during 2010 in England: (a) 1-year net survival (least deprived), (b) 1-year net survival (most deprived), (c) 5-year net survival (least deprived), (d) 5-year net survival (most deprived)

results presented in the first case study, we observe that net survival is consistently lower for the most deprived category across all regions, which becomes more evident at 5 years after diagnosis. Moreover, we notice some variability in net survival by region. Public Health England annually reports an Index of Cancer Survival (Quaresma et al., 2015), where, in previous years, a clear north-south gradient in survival was reported (Index of cancer survival for Clinical Commissioning Groups in England). However, this gradient has been consistently narrowing, which is also in line with the results reported in Figure 4. This type of targeted descriptive spatial summaries is crucial for generating hypotheses which may serve as a basis for conducting more in-depth investigations about the factors driving the observed inequalities.

7 | DISCUSSION

We have proposed a unifying framework for excess hazard estimation using a link-based additive model formulation, which allows for a variety of link functions, several types of covariate effects, and for all types of censoring and left truncation. Estimation is based on a carefully constructed efficient and stable penalized likelihood-based algorithm. Under standard conditions for generalized additive models adapted to the excess hazard setting, we have shown consistency and asymptotic normality of the estimators. An intuitive implementation of the proposed methodology is available in the `R` package `GJRM`, including the straightforward extraction of relevant quantities, such as the (sub-)population net survival and associated intervals. This is true for any of the smooths included in the model as well. The simulation study, covering four sample sizes and several DGPs, demonstrated the reliability and flexibility of the proposed model. In particular, we have observed low levels of bias and variance of the point estimates, as well as a good ability to recover the excess hazard shape, while maintaining low computational times with the fitting procedure taking between 1 and 30 s for data sets of up to $n = 5000$ observations.

The two case studies using real population-based cancer data highlight the usability of our methodology, which allows for the inclusion of complex effects to answer challenging research questions. We explored socio-demographic inequalities and geographical disparities in net survival for three of the most common cancer types diagnosed in England. In a wider context, such kind of results are increasingly being used to formulate cancer control strategies and to prioritize cancer control measures (All-Party Parliamentary Group on Cancer, 2017; Department of Health, 2019). Other uses of our methodology include, but are not limited to, the study of long-term trends in net survival to evaluate the effectiveness of national cancer plans after they have been implemented, by assessing their impact on survival (Exarchakou et al., 2018). In addition, the possibility of modelling spatial effects through the MRF approach facilitates the study of geographical disparities in cancer survival for different sets of relevant health geographies. We note that age-standardization techniques, that is re-weighting the estimates of net survival using a standard cancer population age distribution, have not been applied to the net survival estimates presented in the case studies. We emphasize that when the interest lies in comparing levels of survival between different populations or over time within the same population, such techniques can be applied to avoid that comparisons are masked by differences in the age profiles of cancer patients, since for most cancers the cancer-specific hazard is age dependent (Corazziari et al., 2004). This is not to be confused with the use of the standardized age at diagnosis as a covariate in the model specification, which is indeed done in the case studies.

Our contribution to the relative survival setting brings new research opportunities, for instance: (a) extending the proposed methodology to model multivariate survival data; a potential direction for such development consists of following the copula link-based model proposed in Marra and Radice (2020); (b) the additive decomposition of the total (or overall) hazard in the relative survival framework, relies on the assumption that the two competing risks (associated with cancer and with all other causes of death) remain independent during the entire follow-up period. Copula functions could be used to relax this assumption; (c) developing formal model selection tools for additive excess hazard regression models (Maringe et al., 2019; Rossell & Rubio, 2022); and (d) extending the applicability of our methodology by implementing additional quantities of interest for cancer research (Belot et al., 2019) in the `GJRM` package.

ACKNOWLEDGEMENTS

AE was partly supported by the Windsor Fellowship DeepMind Computer Science Scholarships and by the UCL Departmental Teaching Assistantship Scholarship. MQ was funded through the Cancer Research UK Population Research Committee Funding Scheme: Cancer Research UK Population Research Committee—Programme Award C7923/A29018. GM and RR were supported by the EPSRC grant EP/T033061/1 during the revision of the work which followed the first submission. Finally, the authors thank the two anonymous reviewers, the associate editor and the editor for their well thought out suggestions which helped us improve and clarify several aspects of the article.

ORCID

Giampiero Marra https://orcid.org/0000-0002-9010-2646

REFERENCES

All-Party Parliamentary Group on Cancer. (2017) All party parliamentary group on cancer inquiry: progress of the England cancer strategy: delivering outcomes by 2020?

Belot, A., Ndiaye, A., Luque-Fernandez, M., Kipourou, D., Maringe, C., Rubio, F. et al. (2019) Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data. *Clinical Epidemiology*, 11, 53.

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

Coleman, M. (2014) Cancer survival: global surveillance will stimulate health policy and improve equity. *The Lancet*, 383(9916), 564–573.

Corazziari, I., Quinn, M. & Capocaccia, R. (2004) Standard cancer patient population for age standardising survival ratios. *European Journal of Cancer*, 40, 2307–2316.

Cox, D. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–202.

Cramb, S., Mengersen, K., Lambert, P., Ryan, L. & Baade, P. (2016) A flexible parametric approach to examining spatial variation in relative survival. *Statistics in Medicine*, 35(29), 5448–5463.

Department of Health. (2011) Improving outcomes: a strategy for cancer. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/213785/dh_123394.pdf

Department of Health. (2019) NHS long term plan for cancer. https://www.longtermplan.nhs.uk/

Estève, J., Benhamou, E., Croasdale, M. & Raymond, L. (1990) Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*, 9(5), 529–538.

Exarchakou, A., Rachet, B., Belot, A., Maringe, C. & Coleman, M. (2018) Impact of national cancer policies on cancer survival trends and socioeconomic inequalities in England, 1996–2013: population based study. *BMJ*, 360:k764.

Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., Remontet, L. & Challenges in the Estimation of Net Survival Working Survival Group. (2019) Multi-dimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5), 1233–1257.

Hjort, N.L. (1992) On inference in parametric survival data models. *International Statistical Review*, 60(3), 355–387.

Lambert, P. & Royston, P. (2009) Further development of flexible parametric models for survival analysis. *The Stata Journal*, 9(2), 265–290.

Leitenstorfer, F. & Tutz, G. (2007) Generalized monotonic regression based on B-splines with an application to air pollution data. *Biostatistics*, 8(3), 654–673.

Liu, X.-R., Pawitan, Y. & Clements, M. (2018) Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5), 1531–1546.

Maringe, C., Belot, A., Rubio, F. & Rachet, B. (2019) Comparison of model-building strategies for excess hazard regression models in the context of cancer epidemiology. *BMC Medical Research Methodology*, 19(1), 1–18.

Marra, G. & Radice, R. (2020) Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, 115(530), 886–895.

Marra, G. & Radice, R. (2021) GJRM: generalised joint regression modelling. R package version 0.2-4.

Office for National Statistics, Newport, UK. (2011) Cancer registration statistics, England: 2011.

Pavlič, K. & Pohar-Perme, M. (2019) Using pseudo-observations for estimation in relative survival. *Biostatistics*, 20(3), 384–399.

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

Pohar-Perme, M., Henderson, R. & Stare, J. (2009) An approach to estimation in relative survival regression. *Biostatistics*, 10(1), 136–146.

Pohar-Perme, M., Esteve, J. & Rachet, B. (2016) Analysing population-based cancer survival—settling the controversies. *BMC Cancer*, 16(1), 1–8.

Pya, N. & Wood, S. (2015) Shape constrained additive models. *Statistics and Computing*, 25(3), 543–559.

Quaresma, M., Coleman, M. & Rachet, B. (2015) 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971.2011: a population-based study. *The Lancet*, 385(9974), 1206–1218.

Quaresma, M., Carpenter, J. & Rachet, B. (2019) Flexible Bayesian excess hazard models using low-rank thin plate splines. *Statistical Methods in Medical Research*, 29(6), 1700–1714.

Rachet, B., Maringe, C., Woods, L., Ellis, L., Spika, D. & Allemani, C. (2015) Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health*, 15, 1240.

Reid, N. (1994) A conversation with Sir David Cox. *Statistical Science*, 9(3), 439–455.

Rossell, D. & Rubio, F. (2022) Additive Bayesian variable selection under censoring and misspecification. *Statistical Science*, in press.

Rubio, F., Remontet, L., Jewell, N. & Belot, A. (2019) On a general structure for hazard-based regression models: an application to population-based cancer research. *Statistical Methods in Medical Research*, 28, 2404–2417.

Rubio, F., Rachet, B., Giorgi, B., Maringe, C. & Belot, A. (2021) On models for the estimation of the excess mortality hazard in case of insufficiently stratified life tables. *Biostatistics*, 22(1), 51–67.

Rue, H. & Held, L. (2005) *Gaussian Markov random fields: theory and applications*. Boca Raton, FL: Chapman & Hall/CRC.

Wood, S.N. (2017) *Generalized additive models: an introduction with R*, 2nd edition, London: Chapman & Hall/CRC.

Wood, S.N., Pya, N. & Säfken, B. (2016) Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516), 1548–1563.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Eletti, A., Marra, G., Quaresma, M., Radice, R. & Rubio, F.J. (2022) A unifying framework for flexible excess hazard modelling with applications in cancer epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1–19. Available from: https://doi.org/10.1111/rssc.12566