Flexible Bayesian Excess Hazard models using Low-Rank Thin Plate splines

Journal Title
XX(X):1–22

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Manuela Quaresma[1], James Carpenter[1,2] and Bernard Rachet[1]

Abstract

Excess hazard models became the preferred modelling tool in population-based cancer survival research. In this setting, the model is commonly formulated as the additive decomposition of the overall hazard into two components: the excess hazard due to the cancer of interest and the population hazard due to all other causes of death. We introduce a flexible Bayesian regression model for the log-excess hazard where the baseline log-excess hazard and any non-linear effects of covariates are modelled using Low Rank Thin Plate splines. Using this type of splines will ensure that the log-likelihood function retains tractability not requiring numerical integration. We demonstrate how to derive posterior distributions for the excess hazard and for net survival, a population-level measure of cancer survival that can be derived from excess hazard models. We illustrate the proposed model using survival data for patients diagnosed with colon cancer during 2009 in London, England.

Keywords

Population-based, cancer, survival, excess hazard, Bayesian, flexible, low rank thin plate splines

Introduction

Regression models for the excess hazard became the preferred modelling tool for cancer survival research using population-based data [1–3]. In the absence of reliable recording of the cause of death for each cancer patient, these models conveniently allow to filter out the hazard due to other causes of death, whilst focusing inferences on the excess hazard only due to the cancer of interest. In this setting, the model

[1]London School of Hygiene & Tropical Medicine, Faculty of Epidemiology & Population Health, London, UK. [2]London Hub for Trials Methodology Research, MRC Clinical Trials Unit at UCL, London, UK.

Corresponding author:
Manuela Quaresma, London School of Hygiene & Tropical Medicine, Faculty of Epidemiology & Population Health, Keppel Street, WC1E7HT London, UK
Email: Manuela.Quaresma@lshtm.ac.uk

Prepared using sagej.cls *[Version: 2017/01/17 v1.20]*

is formulated as the additive decomposition of the total hazard into two components: the hazard due to the cancer (the main quantity of interest, also designated as the excess hazard), and the hazard due to all other causes of death, derived from population life tables (also known as background mortality or expected hazard). This set-up allows inequalities in cancer survival to be investigated by looking at the effect of multiple prognostic factors on the form of the excess hazard function or by deriving excess hazard ratios for different sets of characteristics of the population. Model parameter estimates can also be used to derive net survival, both at the individual-level and at the population-level, measuring the survival that can be attributed only to the cancer of interest after accounting for all other causes of death[4]. Net survival estimates can therefore be compared even if the expected hazard differs widely between sub-populations of patients[5]. In their seminal paper, Estève et al.[1] introduced the first regression model for the excess hazard based on the full-likelihood specification using individual survival time data. In its original formulation, the model was proposed on the log-excess hazard scale with the baseline log-excess hazard modelled as a piecewise constant step function, and allowing proportional effects of covariates and linear effects for continuous variables to be modelled. Proposed extensions to this model mainly relaxed the non-proportionality and non-linearity assumptions for covariates and interaction terms, and non-linearity for the baseline excess hazard, by modelling these terms with highly flexible functions such as the commonly used B-splines or restricted cubic splines[2,6]. The tradeoff for the increased model flexibility obtained with the use of splines, is the added complexity to the likelihood function, that requires advanced numerical integration techniques such as the Cavalieri-Simpson integration[2] or the Gaussian quadrature[6] to evaluate the cumulative hazard integral, which will no longer be a tractable function with a closed-form solution.

This applies regardless of the framework of inference, whether frequentist or Bayesian, although inferences for excess hazard models have mainly been based on the frequentist maximisation of the likelihood function. Very few options are available for inferences within the Bayesian framework[7–10], in particular none describing the process of deriving a posterior distribution of net survival.

The purpose of this article is to introduce a flexible Bayesian regression model for the log-excess hazard, based on individual-level data, with the following characteristics: a) the baseline log-excess hazard is modelled using a flexible function; b) the log-likelihood function retains tractability so that numerical integration is not required; c) the model can accommodate a variety of covariate effects: linear and non-linear (also modelled using a flexible function), proportional and non-proportional; d) one can derive a posterior distribution for the excess hazard, excess hazard ratios and net survival; e) the model can be easily extended to include random effects and hierarchical data structures; f) inference can be done within the Bayesian framework; g) and the model can be implemented using most Bayesian open-source software.

Section 2 specifies the likelihood for the log-excess hazard model, introduces the formulation of the flexible functions used in this article, and describes the Bayesian inference procedure, including the steps to obtain a Bayesian posterior distribution for the excess hazard function, excess hazard ratios and net survival. Section 3 provides an example of application of the proposed model based on the survival time data of patients diagnosed with colon cancer during 2009 in London. Section 4 presents some concluding remarks, discusses the limitations of our study, and proposes further extensions to this work.

Methods

Likelihood formulation for the Excess Hazard Model (EHM)

Let $(t_i, \mathbf{x}_i, \delta_i)$, i=1,..., n, $t_i > 0$, denote a set of n time to event observations, measured from the date of diagnosis of a cancer until the occurrence of death, with covariates $\mathbf{x}_i$ and vital status indicator δ_i (δ_i=0 if censored, δ_i=1 if death occurred). The likelihood function of the full vector of parameters of interest $\boldsymbol{\theta}$ can be written in generic terms as

$$L(\theta) = \prod_{i=1}^{n} h(t_i, \mathbf{x}_i, \theta)^{\delta_i}.S(t_i, \mathbf{x}_i, \theta) \tag{1}$$

where $h(t_i, \mathbf{x}_i, \theta)$ is the hazard function and $S(t_i, \mathbf{x}_i, \theta)$ is the survivor function. Delayed entry or left-truncation of observations, can be accommodated in the likelihood by including an additional term, $S(t_d, \mathbf{x}_i, \theta)$, representing the survivor function for a pre-specified truncation time $t_d \geq 0$, as

$$L(\theta) = \prod_{i=1}^{n} \frac{h(t_i, \mathbf{x}_i, \theta)^{\delta_i}.S(t_i, \mathbf{x}_i, \theta)}{S(t_d, \mathbf{x}_i, \theta)} \tag{2}$$

If $t_d > 0$ then $S(t_d, \mathbf{x}_i, \theta) \neq 1$, and the likelihood in equation (2) allows delayed entry of observations[22], enabling study designs such as period analysis to be incorporated into the framework[23]. If $t_d = 0$ then $S(t_d, \mathbf{x}_i, \theta) = 1$, and the likelihood assumes no delayed entry, simplifying to equation (1). For the purpose of this article, the likelihood in equation (1) is used from here onwards, assuming no delayed entry of observations, but the likelihood in equation (2) could be used equivalently in what follows below.

Considering only the individual contribution of observation t_i, the log-likelihood can be written as

$$logL(\theta) = \delta_i.log(h(t_i, \mathbf{x}_i, \theta)) + log(S(t_i, \mathbf{x}_i, \theta)) \tag{3}$$

Using the following relationship between the survival function and the cumulative hazard function ($H(t_i, \mathbf{x}_i, \theta)$)[11]:

$$log(S(t_i, \mathbf{x}_i, \theta)) = -H(t_i, \mathbf{x}_i, \theta) = -\int_0^{t_i} h(u, \mathbf{x}_i, \theta)du \tag{4}$$

and replacing equation (4) into equation (3), the contribution of observation t_i to the log-likelihood can be rearranged as

$$logL(\theta) = \delta_i.log(h(t_i, \mathbf{x}_i, \theta)) - \int_0^{t_i} h(u, \mathbf{x}_i, \theta)du \tag{5}$$

An excess hazard model assumes the additive decomposition of the overall hazard, $h(t_i, \mathbf{x}_i, \theta)$, into two components:

$$h(t_i, \mathbf{x}_i, \theta) = h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i) \tag{6}$$

where, $h_E(t_i, \mathbf{x}_i, \theta)$ is the excess hazard function due to the cancer of interest for an observation t_i with $\mathbf{x}_i$ a vector of observed covariates and θ a set of parameters. The second component, $h_P(a_i + t_i, \mathbf{z}_i)$, is the general population hazard function for an observation t_i, evaluated at the attained age at death (or age at censoring): $a_i + t_i$, with a_i the age at diagnosis and $\mathbf{z}_i$ ($\mathbf{z}_i \in \mathbf{x}_i$) a subvector of covariates for which the

population hazard is defined. The population hazard, also known as background mortality, represents the hazard due to all other causes of death than the cancer of interest. It is assumed to be a known quantity, taken as the age-specific mortality rates from existing population life tables, stratified as finely as possible according to a subset of covariates $\mathbf{z}_i$. This subset of covariates usually contains less covariates than the complete set of covariates available for the cohort of cancer patients, possibly including, in addition to age at death (or censoring), gender and calendar year, socio-economic status, ethnicity or region of residence [12].

Replacing equation (6) into equation (5), the log-likelihood for an excess hazard model can be written entirely as a function of the excess hazard and the population hazard:

$$logL(\theta) = \delta_i.log[h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i)] - \int_0^{t_i} h_E(u, \mathbf{x}_i, \theta)du - \int_0^{a_i+t_i} h_P(u, \mathbf{z}_i)du \quad (7)$$

Given that the population hazard $h_P(a_i + t_i, \mathbf{z}_i)$ is a constant, the last integral in equation (7) does not depend on any parameters and thus can be dropped from the log-likelihood, which can be rewritten (up to this constant) as:

$$logL(\theta) \propto \delta_i.log[h_E(t_i, \mathbf{x}_i, \theta) + h_P(a_i + t_i, \mathbf{z}_i)] - \int_0^{t_i} h_E(u, \mathbf{x}_i, \theta)du \quad (8)$$

Modelling the Log-Excess Hazard function

Equation (8) specifies the log-likelihood for a generic excess hazard model. Inferences can be made by specifying an appropriate model for the excess hazard function ($h_E(t, \mathbf{x})$), which we here assume to have a multiplicative effect of the covariates on the baseline excess hazard. It can be written as,

$$h_E(t, \mathbf{x}) = h_{E_0}(t).exp(\beta.\mathbf{x}) \quad (9)$$

where, $h_{E_0}(t)$ is the baseline excess hazard; and $\mathbf{x} = (x_1, x_2, x_3, \ldots)$ a vector of observed covariates and $\beta = (\beta_1, \beta_2, \beta_3, \ldots)$ the vector of their corresponding parameters. In this article, we propose a model for the logarithm of the excess hazard function, that can accommodate several types of covariate effects. Taking the logarithm of equation (9), we can write, in generic terms, a model for the logarithm of the excess hazard as

$$log(h_E(t, \mathbf{x})) = log(h_{E_0}(t)) + \beta_1.x_1 + g_1(x_2) + g_2(t).x_3 + \ldots \quad (10)$$

where, $log(h_{E_0}(t))$ is now the baseline log-excess hazard function; β_1 is a linear and proportional effect on the log-excess hazard of covariate x_1; $g_1(x_2)$ is a non-linear and proportional effect of a continuous covariate x_2; $g_2(t)$ is a non-proportional (i.e. time-dependent) effect of a covariate x_3.

We choose different constructs of low-rank thin-plate (LRTP) splines to model the baseline log-excess hazard any non-linear effects, and to accommodate time-dependent effects. These first-order polynomials are a penalised type of radial basis splines [13], that have been discussed by several authors for their simple yet flexible nature, providing a good alternative to other spline constructs, such as B-splines and truncated basis splines. In particular, LRTP splines exhibit fast Markov Chain Monte Carlo (MCMC) convergence properties and conveniently result in tractable likelihood functions [13,14]. Murray et al. [15] have introduced a

unified framework for flexible, fully Bayesian analysis of overall survival using LRTP splines, providing a detailed description of their formulation ([15]: Appendix-A), and also making available user-friendly code for easy practical implementation. We follow this spline implementation in the work presented here, and for completeness, in the next three sections, we provide a brief enunciation of the LRTP splines we use to model the different components of the excess hazard model, but do not go into detail about their implementation.

Modelling the baseline log-excess hazard We start by specifying a partition of the follow-up time range as $0 = \tilde{t}_0 < \tilde{t}_1 < \ldots < \tilde{t}_k = \infty$, and following the model formulation published in Murray et al.[15], we define the model for the baseline log-excess hazard as,

$$log(h_{E_0}(t;\alpha^*)) = \alpha_0^* + \alpha_1^* t + \sum_{k=2}^{K} \alpha_k^* (|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \tag{11}$$

where $\alpha^* = (\alpha_0^*, \alpha_1^*, \ldots, \alpha_K^*)^{'}$ is the set of spline parameters. Under equation (11), the cumulative excess hazard takes the expression,

$$H_{E_0}(t;\alpha^*) = \sum_{k=1}^{K} \frac{h_{E_0}(s_k;\alpha^*)[1 - e^{-(s_k - \tilde{t}_{k-1})(u_{k,K}^{'}.\alpha_{(-1)}^*)}]}{u_{k,K}^{'}.\alpha_{(-1)}^*} \tag{12}$$

where s_k=max(min(t, $\tilde{t}_k$), $\tilde{t}_{k-1}$), $\alpha_{(-1)}^* = (\alpha_1^*, \ldots, \alpha_K^*)^{'}$, $u_{k,K}^{'} = (1_k^{'}, -1_{K-k}^{'})$, for $k = 1, \ldots, K$ with $1_k^{'}$ a k-dimensional vector of ones. Implementation of this spline involves a series of transformations to the spline parameters α^*, as well as constructing a time design matrix and a penalty transformation matrix, as detailed by Crainiceanu et al.[14] and Murray et al.[15], so that the baseline log-excess hazard can be rewritten in terms of these transformed components.

Modelling a non-linear effect of a continuous covariate We model any non-linear effect of a generic continuous covariate x, as a smooth effect using a cubic LRTP spline defined as,

$$g(x;\beta^*) = \beta_1^*(x - \overline{x}) + \sum_{j=2}^{J} \beta_j^* (|x - \widetilde{x}_{j-1}|^3 - |\overline{x} - \widetilde{x}_{j-1}|^3) \tag{13}$$

where, $\beta^* = (\beta_1^*, \beta_2^*, \ldots, \beta_J^*)$ is a set of spline parameters, $\overline{x}$ is the sample mean of covariate x, and $(\widetilde{x}_1 < \widetilde{x}_2 < \ldots < \widetilde{x}_J)$ is a partition of the covariate's support range. Similarly to the model specification for the baseline log-excess hazard, implementation will require one-to-one transformations to reparametrise (13) in terms of β^* [15].

Incorporating a non-proportional (time-dependent) effect of a covariate To incorporate a time-dependent effect of a generic covariate x, we use the same time partition as used for the baseline log-excess hazard as in equation (11), and define,

$$log(h_E(t|x;\alpha^*)) = (\alpha_{0,0}^* + \alpha_{1,0}^* x) + (\alpha_{0,1}^* + \alpha_{1,1}^* x)t + \sum_{k=2}^{K} (\alpha_{0,k}^* + \alpha_{1,k}^* x)(|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \tag{14}$$

where $\alpha^* = (\alpha_0^*|\alpha_1^*)$ and $\alpha_q^* = (\alpha_{q,0}^*, \ldots, \alpha_{q,K}^*)'$ for $q = 0, 1$. Similarly to the model defined for the baseline log-excess hazard, implementation involves a series of transformations to the splines parameters α^* to rewrite the time-dependent effect in terms of these transformed components [15].

Prior distributions

For the Bayesian estimation we choose the following prior distributions for the model parameters:

- For the baseline log-excess hazard as specified in equation (11):

$$\begin{gathered} \alpha_0 \sim N(0, 10^4),\ \alpha_1 \sim N(0, 10^4) \\ \alpha_k|\sigma_\alpha \overset{iid}{\sim} N(0, \sigma_\alpha^2), \text{ for k=2}, \ldots, \text{ K and } \sigma_\alpha \sim U(0.01, 100) \end{gathered} \tag{15}$$

- For the parameters of a non-linear effect in equation (13):

$$\begin{gathered} \beta_0 \sim N(0, 10^4) \\ \beta_k|\sigma_\beta \overset{iid}{\sim} N(0, \sigma_\beta^2) \text{ , for k=2, } \ldots \text{ , K and } \sigma_\beta \sim U(0.01, 100) \end{gathered} \tag{16}$$

- And, for the parameters of a time-dependent effect as in equation (14):

$$\begin{gathered} \alpha_{q,0} \sim N(0, 10^4) \text{ , } \alpha_{q,1} \sim N(0, 10^4) \text{ for q=0, 1} \\ \alpha_{q,k}|\sigma_{q,\alpha} \overset{iid}{\sim} N(0, \sigma_{q,\alpha}^2) \text{ for k=2, } \ldots \text{ , K and } \sigma_{q,\alpha} \sim U(0.01, 100) \text{ for q=0, 1} \end{gathered} \tag{17}$$

Measures of interest: excess hazard and net survival

In addition to deriving excess hazard functions and excess hazard ratios for different sets of characteristics of the population, another main quantity of interest that can be derived from an excess hazard model is net survival. Net survival measures the survival in a cohort of cancer patients while considering that the patients can only die from the cancer of interest. A common assumption made when estimating net survival is that the censoring process is non-informative, i.e. the censoring process is independent from the one that generates the events. The process becomes informative when a variable influences both mortality hazards (the cancer-specific and the other-causes mortality hazard), leading to biased estimates of net survival. For example, older patients are more likely to be censored, because of other causes of death, than younger patients, making the censoring process informative. It has been shown that in order to obtain an unbiased estimate of net survival from an excess hazard regression model, the variables that define the population-life tables (from which the other-cause mortality is obtained), and that can influence the censoring process, should be included in the excess hazard model formulation, even if they are not the main focus of the analysis [16]. In population-based cancer research, one of the main variables known to influence the censoring process is age at diagnosis, and thus it is advisable to include it in all the log-excess hazard model formulations. It is also advisable to include other variables in the model formulation, such as socio-economic status or region of residence, if life-tables stratified by these variables are available for the population being studied.

Bayesian estimation

After setting up a model for the log-excess hazard, possibly using a combination of several covariate effects modelled using LRTP splines, as specified in the previous sections, the resulting log-likelihood function retains tractability, and thus numerical integration techniques are not needed during the estimation process. Markov Chain Monte Carlo (MCMC) techniques are used to sample from the posterior distributions of all the model parameters. After model convergence has been assessed by inspecting trace and density plots for each parameter, the saved parameter samples are used in a post-estimation procedure to derive posterior distributions of the quantities of interest that can be obtained from excess hazard models. We present post-estimation set-ups to derive posterior distributions of: i) excess hazards, ii) excess hazard ratios for different combinations of population characteristics and iii) net survival for the whole population and for sub-groups of the population.

i) Deriving posterior distributions of excess hazards

Figure 1 shows schematically the post-estimation set-up to derive posterior distributions of excess hazards for different combinations of population characteristics.

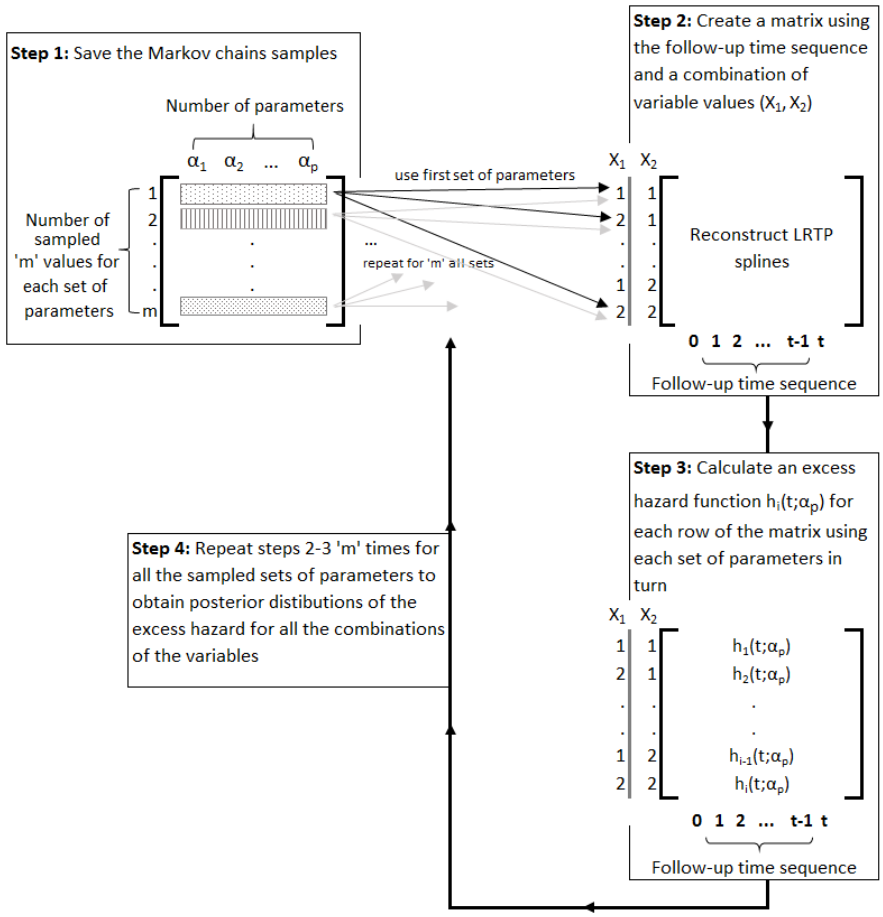

Figure 1. Step-by-step set-up to derive the posterior distributions of excess hazards.

The procedure can be summarised in the following steps:

Step 1 Create a matrix that saves for each parameter (say generically $\alpha_i, i = 1, \ldots, p$) the number of sampled Markov chain values (say 'm') from their corresponding posterior distributions.

Step 2 Create a follow-up time sequence within the observed range of follow-up time (0,...,max(t)), and create a matrix containing this time sequence and a combination of values from the variables

entered in the model, chosen within the observed range of values for each variable (e.g. variables X1 and X2 in Figure 1); Re-construct the LRTP splines as defined in the model using the values in the matrix, both for the baseline log-excess hazard and all the covariate effects modelled with LRTP splines.

Step 3 Use the first set of the 'm' sampled parameters to estimate an excess hazard function for each combination of variables in the matrix using the follow-up time sequence.

Step 4 Repeat steps *2-3* 'm' times for all the sets of sampled parameters (in turn) to obtain posterior distributions of the excess hazard functions for all the combinations of variable characteristics.

Step 5 Summarise the posterior distributions of the excess hazards using the posterior means, 95% credible intervals and other relevant quantiles.

ii) Deriving posterior distributions for excess hazard ratios

Figure 2 shows the post-estimation set-up for deriving posterior distributions of excess hazards ratios.

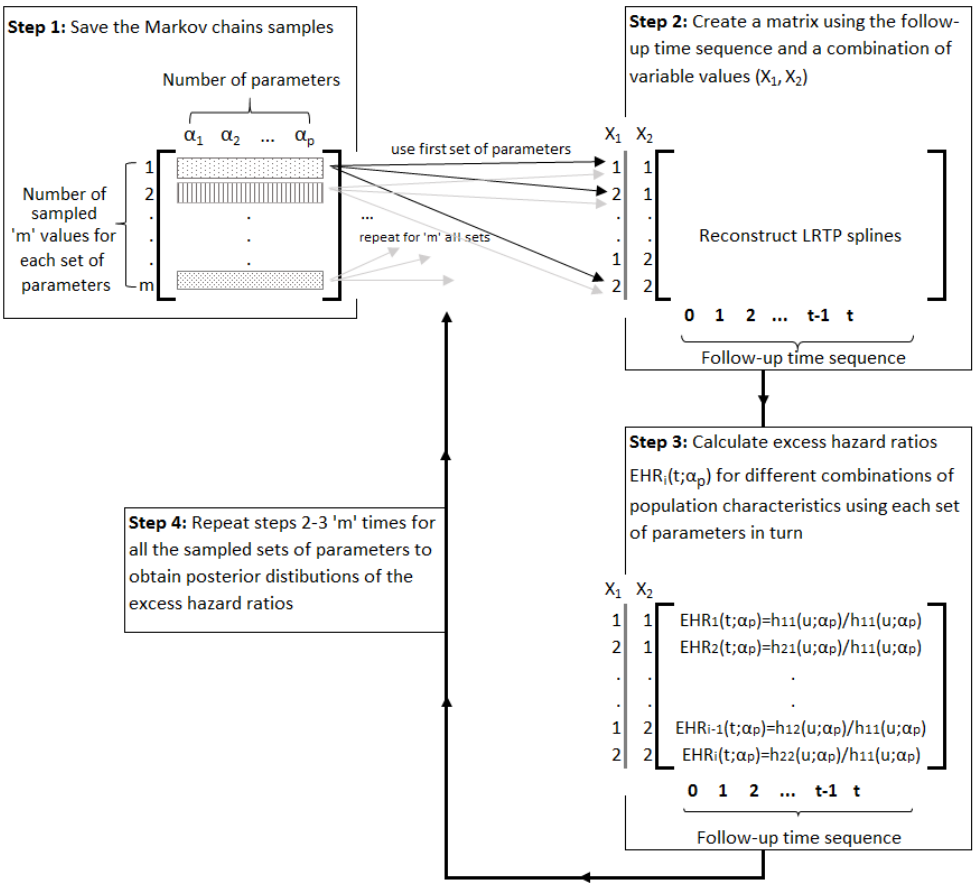

Figure 2. Step-by-step set-up to derive the posterior distributions of excess hazard ratios.

Similarly to the procedure defined in Figure 1 for the excess hazards, the procedure to derive posterior distributions of excess hazard ratios can be summarised as:

Step 1 Same as **Step 1** from the set-up in Figure 1.

Step 2 Same as **Step 2** from the set-up in Figure 1.

Step 3 Use the first set of the 'm' sampled parameters to estimate excess hazard ratios for different combinations of variables in the matrix using the follow-up time sequence.

Step 4 Repeat steps *2-3* 'm' times for all the sets of sampled parameters (in turn) to obtain posterior distributions of the excess hazard ratios for all the combinations of variables.

Step 5 Summarise the posterior distributions of the excess hazards ratios using the posterior means, 95% credible intervals and other relevant quantiles.

iii) Deriving posterior distributions of net survival

Figure 3 shows the post-estimation set-up for deriving posterior distributions of net survival for the whole population.

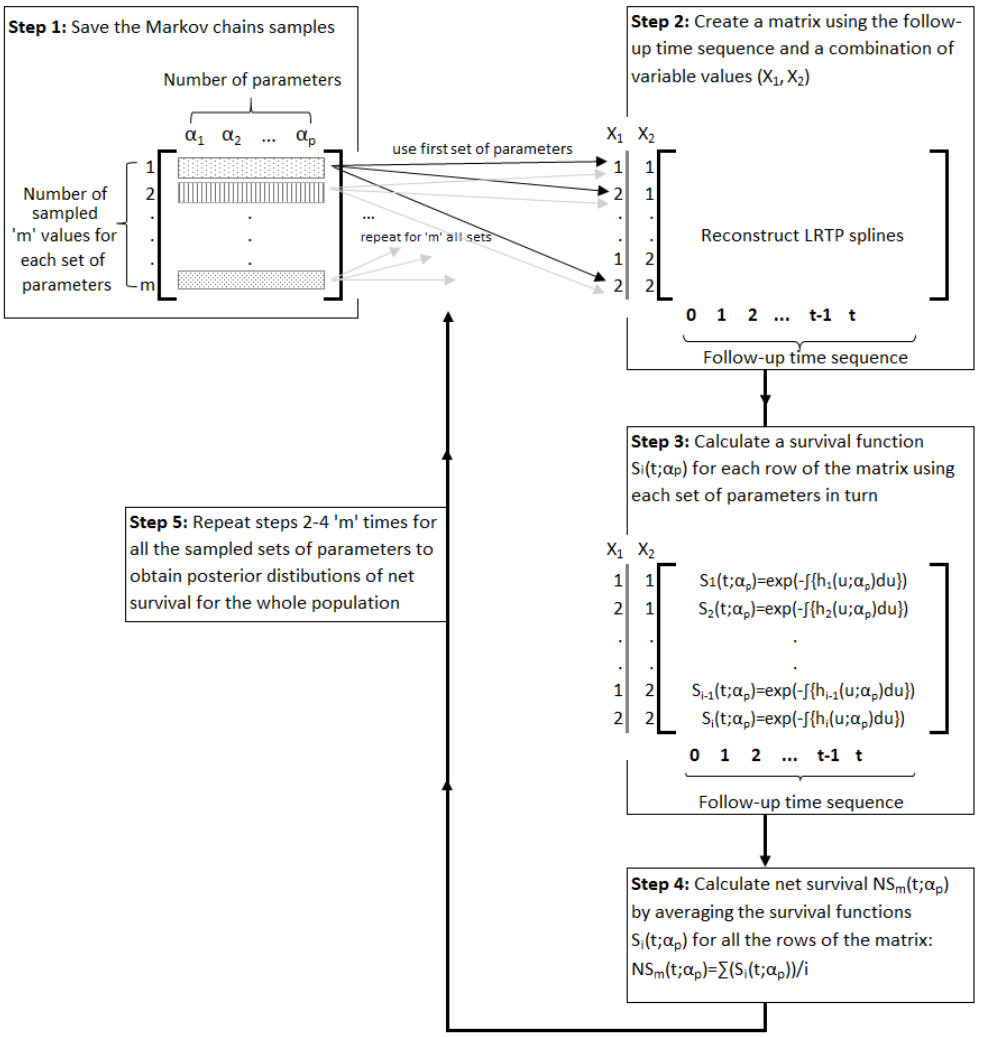

Figure 3. Step-by-step set-up to derive the posterior distributions of net survival for the whole population.

The procedure to derive posterior distributions of net survival can be summarised as:

Step 1 Same as **Step 1** from the set-up in Figure 1.

Step 2 Same as **Step 2** from the set-up in Figure 1.

Step 3 Use the first set of the 'm' sampled parameters to estimate a survival function for each entry of the matrix using the follow-up time sequence.

Step 4 Calculate net survival for the whole population by averaging the survival functions for all the rows of the matrix derived in step *3*.

Step 5 Repeat steps *2-3-4* 'm' times for all the sets of sampled parameters (in turn) to obtain posterior distributions of net survival for the whole population.

Step 6 Summarise the posterior distributions of net survival using the posterior means, 95% credible intervals and other relevant quantiles.

The implementation above will provide posterior distributions of net survival for the whole population. We can also derive posterior distributions of net survival by sub-groups of the population, continuing from **step 3** in Figure 3 and averaging the survival functions within each sub-group of the population (e.g. by sub-groups of the variable X_2 as shown by **step 4** in Figure 4).

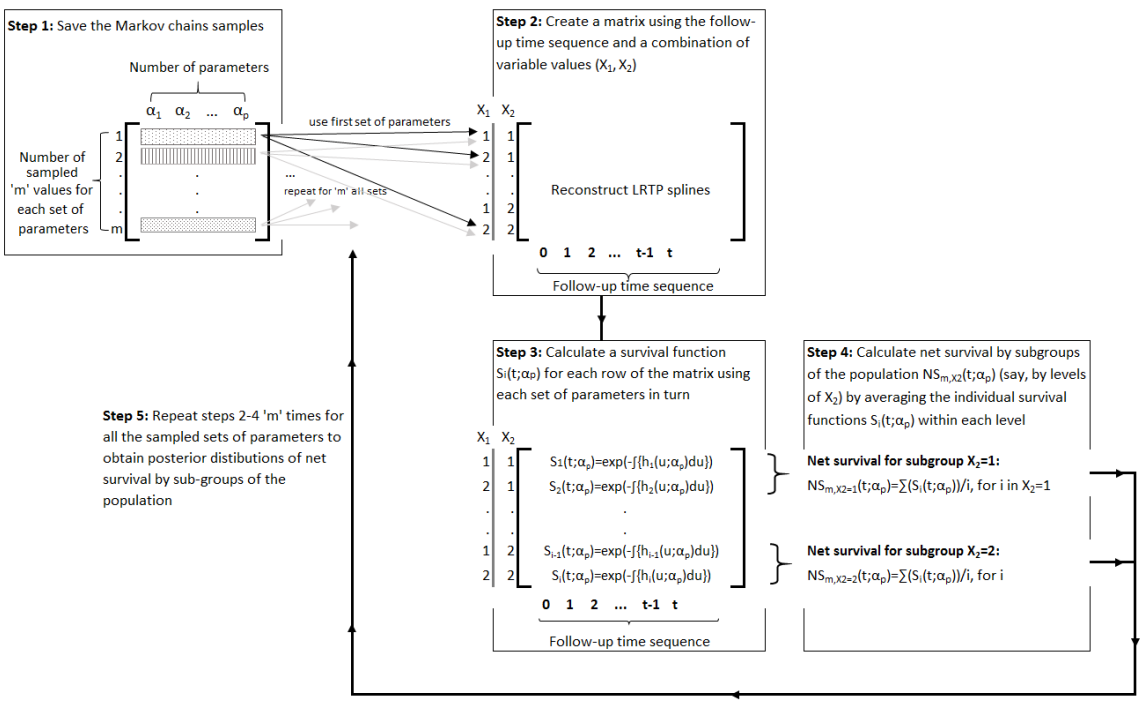

Figure 4. Set-up to derive the posterior distributions of net survival by sub-groups of the population.

Illustration using population-based cancer data

We illustrate the use of the proposed model using data obtained from the National Cancer Registry at the Office for National Statistics (ONS) for all adult men (aged 15-99 years) diagnosed with a first, primary, invasive malignancy of the colon during 2009 in London, England. All patients were followed-up to update their vital status up to six years after diagnosis, until the 31 December 2015. The data variables available for this analysis were: full dates of diagnosis, last follow-up and death, vital-status indicator (dead or censored as alive at the end of follow-up), age at diagnosis (recorded as a continuous variable) and deprivation categories (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[17]. Background mortality rates were obtained for each cancer patient from population life tables for England defined for each calendar year in 2009-2015, and stratified by single year of age, sex, deprivation category and region of residence.

Descriptive statistics of the data were performed using the *RStudio* software (version 1.0.153)[18]. Bayesian inferences were also performed in *RStudio* using the JAGS MCMC[19] program accessed via the *R* package '*R2JAGS*'. R code exemplifying the implementation of the model presented in this illustration is available on the webpage of the *Cancer Survival Group*: https://csg.lshtm.ac.uk/tools-analysis/.

The data comprised 1,140 patients. Death was observed for 628 patients (55.1%) over the maximum follow-up period of 5.99 years. Survival time was measured from the date of diagnosis until the date of death or the date of last follow-up. The overall median follow-up time was 3.7 years with standard deviation SD=2.29 years. For patients that died, the median survival time was 0.84 years and for censored patients the median survival time was 5.4 years. The mean age at diagnosis was 70.6 years (SD=13.24 years), and the 25%, 50% and 75% quintiles of the age distribution were 63.2, 72.4 and 80.6 years, respectively. Within deprivation categories, patients were distributed as: 174 (15%) patients in the least deprived category, 207 (18%) patients in the 2nd deprivation category, 223 (20%) patients in the 3rd category, 273 (24%) patients in the 4th category, and 263 (23%) patients in the most deprived category.

A model was set-up for the log-excess hazard including age at diagnosis (A) and deprivation quintile (dep) as main effect covariates. Four partitions (K=4) of the observed follow-up time (t) were chosen at the 25%, 50% and 75% percentiles of the events (death) times at $\widetilde{t}$=(0, 0.18, 0.84, 2.26, 6) years. The model can be written as:

$$\begin{aligned} log(h_E(t|\alpha;\beta;\gamma)) = {} & (\alpha_{0,0} + \alpha_{1,0}A) + (\alpha_{0,1} + \alpha_{1,1}A)t \\ & + \sum_{k=2}^{K}(\alpha_{0,k} + \alpha_{1,k}A)(|t - \tilde{t}_{k-1}| - |\tilde{t}_{k-1}|) \qquad \text{[part 1]} \\ & + \beta_1^*(A - \overline{A}) + \sum_{j=2}^{J}\beta_j^*(|A - \widetilde{A}_{j-1}|^3 - |\overline{A} - \widetilde{A}_{j-1}|^3) \qquad \text{[part 2]} \\ & + \gamma * dep \qquad \text{[part 3]} \end{aligned} \tag{18}$$

where, [part 1] formulates the LRTP spline modelling the baseline log-excess hazard, incorporating the time-dependent effect of age at diagnosis using the same follow-up time partition, with parameters α = $(\alpha_0|\alpha_1)$ and $\alpha_q = (\alpha_{q,0}, \ldots, \alpha_{q,K})$ for q=0,1. [part 2] represents the LRTP spline modelling the non-linear (smooth) effect of age at diagnosis using 3 partitions (J=3) of the observed age range at $\widetilde{A}$=(16,

44, 72, 99) years, with parameters β_j, $j = 1, \ldots, J$. $\overline{A}$ represents the mean age at diagnosis. For ease of interpretation, age at diagnosis was centered at age 70. [part 3] formulates the linear and proportional effect of deprivation, with parameter γ. This model has 10 parameters associated with the baseline log-excess hazard formulation, including the time-dependent effect of age at diagnosis, and 4 parameters for the regression parameters (3 for the smooth effect of age at diagnosis and 1 for the effect of deprivation). Prior distributions were specified for these parameters using the priors defined in the Methods section, including 3 hyperpriors for the variance parameters of these priors, adding up to a total of 17 model parameters.

The model was fitted setting up 2 MCMC chains, each with 50,000 iterations, a burn-in period of 5,000 and a thinning of 3 to eliminate any existing autocorrelation among samples within the chains. This resulted in a total of 30,000 sampled values from the posterior distributions of each of the 17 parameters. An examination of the trace and density plots of each parameter's posterior distribution did not indicate any convergence issues for these samples. The 30,000 sampled values from the posterior distributions of each parameter, were saved and then used to implement the post-estimation procedure described in Fig. 1 in order to derive posterior distributions for the excess hazard, excess hazard ratios and net survival. Three 'prediction' sequences were created for follow-up time (monthly time points up to five years of follow-up), age at diagnosis (individual integer ages within the observed age range 16-99 years) and deprivation category (1-5). A multi-dimensional matrix was then created to save the results of the posterior distributions for each of the quantities derived, containing the combination of values of all the 'prediction' sequences, and the number of sampled parameter values (30,000). Before the post-estimation procedure was implemented, the splines modelling the baseline log-excess hazard and the smooth effect of age at diagnosis, were reconstructed using the follow-up time and age 'prediction' sequences, maintaining the same spline specification as in the model.

The estimated posterior distributions were summarised by their respective means and other quantiles of interest, such as the 95% credible intervals. For the purpose of this illustration, we present the results in plots (Figures 5-7) showing the mean of each of the posterior distributions.

Interpretation summary: For this cohort of men diagnosed with colon cancer in 2009 in London, England, the estimated mean posterior distributions suggest that: 1) the excess hazard peaks substantially high, up to the first year after diagnosis, for men over 80 years when compared to patients aged 70 years. Whilst for men aged 50 and 60 years their excess hazard is substantially lower, up to the first year after diagnosis, when compared to men aged 70 years. 2); the excess hazard increases gradually for each unit increase in the deprivation category; 3) The mean posterior net survival for the whole cohort shows a moderate decay of the survival curve, reaching approximately 0.6 (60%) at 5 years after diagnosis.

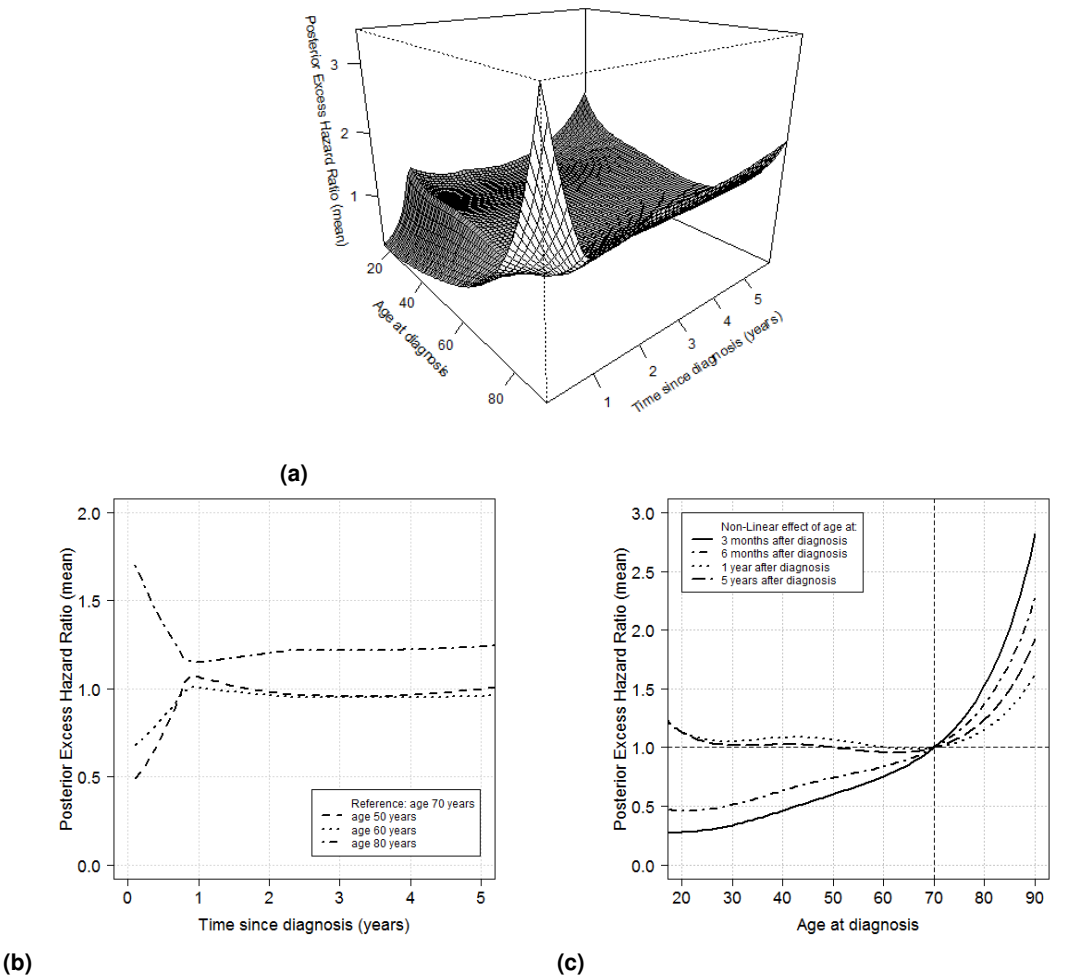

Figure 5. Mean posterior distribution of the excess hazard ratios, showing: (a) a 3-Dimensional representation by age and follow-up time; (b) a slice of the 3-D plot in Fig. 3a) over follow-up time for three ages of diagnosis (50, 60 and 80 years, with 70 years the reference group); (c) a slice of the 3-D plot in Fig. 3a) over age of diagnosis for four follow-up times (3 months, 6 months, 1 year and 5 years after diagnosis).

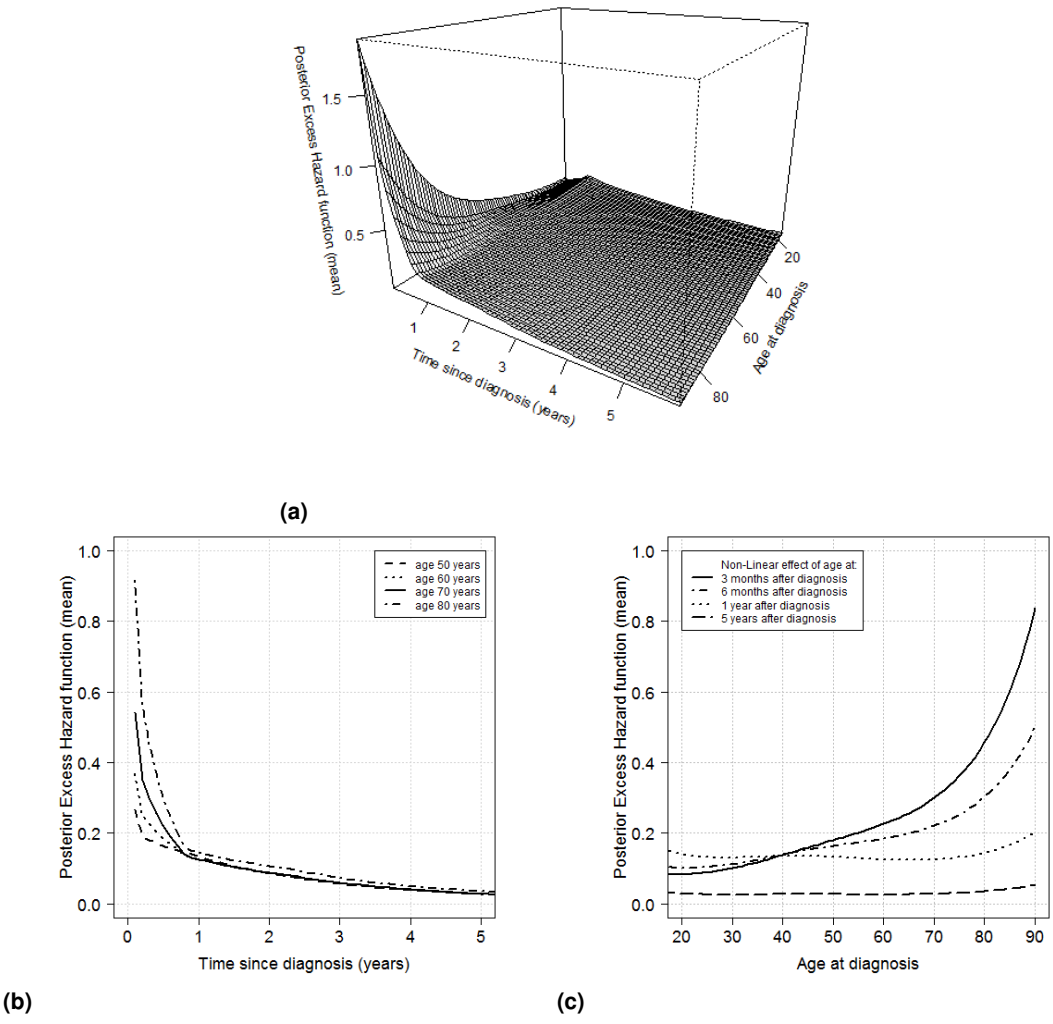

Figure 6. Mean posterior distribution of the excess hazard functions for deprivation category 1 (least deprived patients), showing: (a) a 3-Dimensional representation by age and follow-up time; (b) a slice of the 3-D plot in Fig. 4a) over follow-up time for four age groups (50, 60, 70 and 80 years); and (c) a slice of the 3-D plot in Fig. 4a) over age at diagnosis for four follow-up times (3 months, 6 months, 1 year and 5 years after diagnosis).

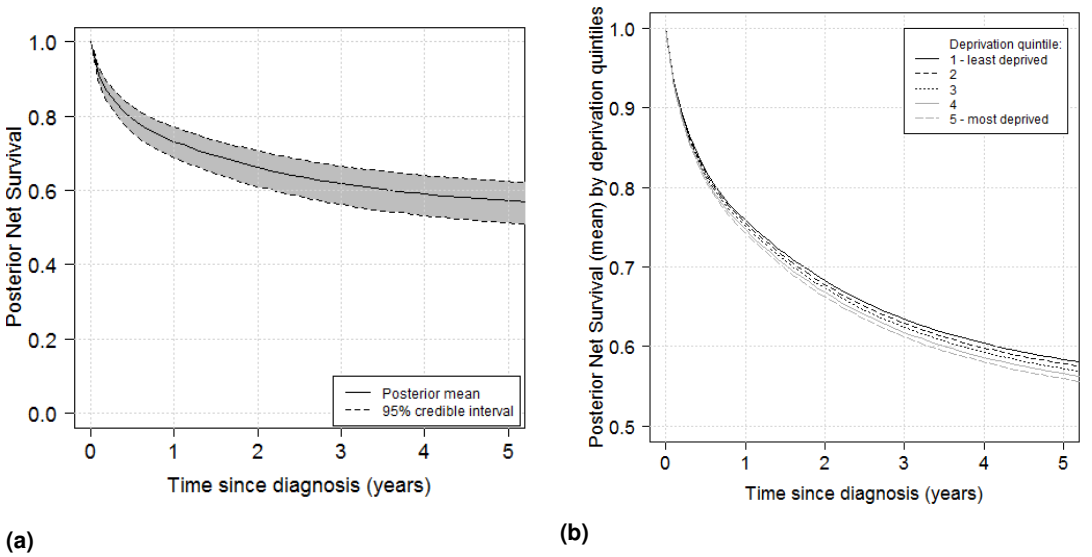

Figure 7. Posterior distribution of net survival, showing: (a) the mean posterior and the 95% credible interval for the whole cohort; and (b) the mean posterior by deprivation category.

Discussion

In this article we introduce a flexible Bayesian regression model for the log-excess hazard, that can be used to investigate inequalities in cancer survival using a range of covariate effects and accommodate different data structures.

Bayesian excess hazard models are very few and none meet our list of criteria set in the Introduction section. Fairly et al.[8] proposed a model to examine spatial variation in prostate cancer survival using Bayesian relative survival smoothing within a Generalised Linear Model formulation. The number of events was assumed to follow a Poisson distribution, and two random effects were included in the model: a spatially structured random effect for local smoothing and an unstructured random effect global smoothing; Hennerfeind[9] proposed a Bayesian geoadditive relative survival model using penalized P-splines to model the log-baseline effect as well as the nonlinear and time-varying effects of covariates. Spatial and normal random effects were also included in the model formulation; Cramb et al.[10] introduced a Bayesian flexible parametric model which extends a frequentist flexible parametric model on the log cumulative excess hazard scale using restricted cubic splines[20] by adding spatially structured random effects.

We choose here to use Low-Rank Thin Plate splines (LRTP splines) to model the various components of the excess hazard model because they offer a reasonable compromise between model flexibility and

likelihood tractability, with a fast MCMC convergence[13,14]. Current inference practice for existing log-excess hazard models are mainly done within the frequentist framework, by maximisation of the log-likelihood, and numerical integration techniques are often needed to solve the integral defining the cumulative hazard when flexible functions, such as restricted cubic splines or B-splines, are used to model the different model components. Incorporating higher-dimensional splines would then require to solve numerically extremely complex likelihood functions. Other existing excess hazard models, that are defined on the log cumulative excess hazard scale, have the advantage of avoiding the use of such numerical integration, because of the resulting tractable cumulative excess hazard, but the interpretation of multiple time-dependent effects can be difficult at times when the excess hazard ratio for one variable depends on the levels of the other variables, even without having defined interaction terms in the model[21].

An additional advantage of using LRTP splines, initially proposed by Murray et al.[15] to model overall hazard, is that their construct is not sensitive to the choice of 'knot' location, as is the case with other splines structures, such as restricted cubic splines or B-splines. Murray et al. advise on the selection of a large number of equally spaced partitions of the follow-up time, so that the resulting model can adequately capture the curvature of the hazard function.

In the analysis of the colon cancer data, we selected several partitions of the follow-up time (between 2 and 20), using a mixture of equally spaced and pre-defined intervals. Models were compared using the Deviance Information Criterion (DIC)[24], and the model presented in the results section (using 4 partitions of the follow-up time) corresponded to the model with the smallest DIC. We found that less partitions (four in our analysis) did adequately capture the shape of the excess hazard function for the cancer analysed, and that partitioning the event times at the percentiles captured well the largest shift in the decay of the function in the first year after diagnosis. The shapes of the baseline excess hazard function and of the age-related function defined in our final model were also very similar to those estimated by the frequentist flexible excess hazard model using higher-dimensional splines[25]. We also observed that using less partitions substantially decreased computation time when fitting these models using MCMC sampling (for example, time was reduced by a quarter when using 4 instead of 20 partitions), as there are less parameters to be sampled. We note that fitting these models can be computationally very expensive, varying from a few hours to a few days, depending on model complexity, the number of MCMC iterations and the size of the matrices generated. Computation time can be reduced by the use of parallel computing, the use of computers with Graphics Processing Units (GPU), or by exploring new advances in accelerated computing such as GPU-accelerated packages[26].

Eliciting informative priors for the model parameters was not within the aim of this study, and we opted to choose vague priors for all the model parameters. In such a scenario, the mean posterior distributions for the parameters and quantities of interest, would be closer to the Maximum Likelihood estimates obtained using a similar model set-up.

A novel component that this article offers, is the implementation of a post-estimation procedure (as described in Fig. 1), to derive posterior distributions for the excess hazard ratios, excess hazard functions and net survival, based on the saved MCMC samples for each parameter. This procedure, as described, derives posterior distributions using a predefined matrix that contains a combination of values of the covariates within the observed range in the data, and it does not use the data for the whole cohort. The estimation of excess hazards and excess hazard ratios is usually made for different sets of characteristics of the cohort, and thus it is easier to construct a matrix to derive these posterior distributions. For net survival, the estimation is made by averaging the individual survival curves, which can be done using one

of two options: 1) use the whole cohort of observed data, estimate a survival curve for each observation (following the same procedure as outlined in Fig. 1), and then average over the whole cohort to obtain an estimate of net survival, or 2) use the matrix, estimate a survival curve for each combination of values of covariates, and then average over these curves to obtain an estimate of net survival. The main advantage of using a matrix over the observed cohort is the reduced computation time, especially when large cohorts are analysed. In addition, when using a fixed covariate structure within the matrix, the results will be internally standardised for those covariates, and thus when comparing net survival by sub-groups of the cohort, this has the advantage that comparability will already be taken into account. For example, if we consider two variables, age at diagnosis and deprivation, and estimate net survival by deprivation category, averaging the individual survival curves within each deprivation group using the whole cohort, if the age distribution within deprivation category is very different, the results will not be comparable. But if we use a matrix with a fixed age structure for all levels of deprivation, the estimated net survival curves will be comparable between deprivation categories.

One of the criteria we set up a priori for the implementation of the proposed model, was that it could be easily extended to include one or more random effects to accommodate clustered data, and incorporate hierarchical data structures. The Bayesian framework lends itself very nicely to specify models with such characteristics. We have extended the model specified in equation (18) to add two random effects: one clustering patients by area of residence, and another clustering patients by treatment center. Although the model implementation was a straight-forward step from the previous model implementation (without random effects), we found some convergency problems when using the open-source MCMC sampler, and a substantial increase in computation time, depending on the size and number of clusters used (results not shown). We propose as further extension to this work, to develop a dedicated MCMC sampler that improves sampling from the parameters' posterior distributions when using these more complex model structures.

In summary, we have shown how a flexible Bayesian model for the log-excess hazard can be used for population-based research, to investigate socio-economic inequalities in cancer survival using a range of covariate effects modelled using LRTP splines. In our experience, we found that using LRTP splines provides a good compromise between the achieved model flexibility and the retained tractability that reduces computational intensity. Although constructing these splines involves many matrix calculations in order to compute the necessary transformations to implement the splines, the user-friendly and modifiable code that has been made available [15] makes the implementation uncomplicated. In particular, we think that the new post-estimation process we propose to derive posterior distributions for net survival and excess hazards will be a very useful tool for cancer researchers in the production of cancer survival statistics with relevance to health policy.

Acknowledgements

The authors wish to acknowledge Dr. Francisco Rubio for the very helpful discussions and encouragements.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The first and last authors wish to thank Cancer Research UK for the funding provided for this research, grant numbers C1336/A11700 and C7923/A18348. The second author wishes to thank the UK Medical Research Council for the funding provided for this research, grant numbers MC_UU_12023/21 and MC_UU/12023/29.

References

1. Estève J, Benhamou E, Croasdale M et al. Relative survival and the estimation of net survival: elements for further discussion. *Statistics in Medicine*. 1990; **9**: 529-538.
2. Remontet L, Bossard N, Belot A et al. An overall strategy based on regression models to estimate relative survival and models to estimate relative survival and model the effects of prognostic factors in cancer survival studies. *Statistics in Medicine*. 2007; **26**: 2214-2228.
3. Lambert PC and Royston P. Further development of flexible parametric models for survival analysis. *Stata Journal*. 2010; **9**: 265-290.
4. Perme MP, Stare J and Estève J. On estimation in relative survival. *Biometrics*. 2012; **68**: 113-120.
5. Perme MP, Estève J and Rachet B. Analysing population-based cancer survival - settling the controversies. *BMC Cancer*. 2016; **16**.
6. Crowther MJ and Lambert PC. A general framework for parametric survival analysis. *Statistics in Medicine*. 2014; **33**: 5280-5297.
7. Giorgi R, Sahel A, Daures JP et al. A Metropolis within Gibbs sampling in relative survival. *Far East Journal of Theoretical Statistics*. 2005; **16**: 269-284.
8. Fairley L, Forman D, West R et al. Spatial variation in prostate cancer survival in the Northern and Yorkshire region of England using Bayesian relative survival smoothing. *British Journal of Cancer*. 2008; **99**: 1786-1793.
9. Hennerfeind A, Held L and Sauleau EA. A Bayesian analysis of relative cancer survival with geoadditive models. *Statistical Modelling*. 2008; **8**: 117-139.
10. Cramb SM, Mengersen KL, Lambert PC et al. A flexible parametric approach to examining spatial variation in relative survival. *Statistics in Medicine*. 2016; **35**: 5448-5463.
11. Collett D. Modelling Survival Data in Medical Research. Chapman & Hall, 2nd ed. 2003.
12. Rachet B, Maringe C, Woods LM et al. Multivariable flexible modelling for estimating complete, smoothed life tables for sub-national populations. *BMC Public Health*. 2015; **15**: 1240.
13. Ruppert D, Wand MP and Carroll RJ. Semiparametric Regression. *Cambridge Series in Statistical and Probabilistic Mathematics*. 2003.
14. Crainiceanu CM, Ruppert D and Wand MP. Bayesian Analysis for Penalized Spline Regression Using WInBUGS. *Journal of Statistical Software*. 2005; **14**: 1-24.
15. Murray TA, Hobbs BP, Sargent DJ et al. Flexible Bayesian Survival Modeling with Semiparametric Time-Dependent and Shape-Restricted Covariate Effects. *Bayesian Analysis*. 2016; **11(2)**: 381-402.
16. Danieli C, Remontet L, Bossard N et al. Estimating net survival: the importance of allowing for informative censoring. *Statistics in Medicine*. 2012; **31(8)**: 775-86.
17. English indices of deprivation. *Ministry of Housing, Communities and Local Government*. 2011. URL https://www.gov.uk/government/collections/english-indices-of-deprivation.
18. RStudio Team. RStudio: Integrated Development for R. 2015. URL http://www.rstudio.com/.
19. Plummer M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. 2003.

20. Nelson CP, Lambert PC, Squire IB et al. Flexible parametric models for relative survival, with application in coronary heart disease. *Statistics in Medicine*. 2007; **26**: 5486-5498.
21. Royston P and Lambert PC. Flexible Parametric Survival Analysis using Stata: Boyond the Cox Model. *Stata Press*. First edition, 2011.
22. Klein JP and Moeschberger ML. Survival analysis: techniques for censored and truncated data. *Springer*, 2003.
23. Brenner H, Gefeller O and Hakulinen T. Period analysis for up-to-date cancer survival data. Theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer*, 2004; **40**: 326-335.
24. Spiegelhalter DJ, Best N and Carlin BP and van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 2002; **64**(4): 583-639.
25. Charvat H, Remontet L, Bossard N, Roche L, Dejardin O, Rachet B, Launoy G, Belot A and CENSUR Working Survival Group. A multilevel excess hazard model to estimate net survival on hierarchical data allowing for non-linear and non-proportional effects of covariates. *Statistics in Medicine*, 2016; **35**: 3066-84.
26. Draper D and Terenin A. Comment: A brief survey of the current state of play for Bayesian computation in data science at big-data scale. *Brazilian Journal of Probability and Statistic*. 2017; **31(4)**: 686–69.