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

Manuela Quaresma

Manuela.Quaresma@lshtm.ac.uk

Prof. Bernard Rachet and Prof. James Carpenter

Cancer Survival Group
London School of Hygiene and Tropical Medicine
Webpage: http://csg.lshtm.ac.uk and Twitter: @csg_lshtm

October 2019

Aim

- Model that could accommodate more complex data structures
  - hierarchical: cancer patients clustered within health geographies (small areas)
- Log-baseline excess hazard modelled using flexible function
- Include non-linear and non-proportional effects
- Inference within Bayesian framework suitable for these model specifications
- **First step**: introduce flexible Bayesian model for the excess hazard (without '*random effects*')

Relative survival setting

Observed hazard decomposed as

$$h(t; \mathbf{x}) = h_E(t; \mathbf{x}) + h_P(A + t; \mathbf{z})$$

- $t = (t_1, \ldots, t_n)$ set of event times, and A=age at diagnosis
- $x = (x_1, \ldots, x_p)$ set of covariates: age, gender, deprivation, comorbidities, tumour stage,$\ldots$, and $\mathbf{z} \subset \mathbf{x}$
- $h_E(t; \mathbf{x})$ - hazard due to cancer: **excess hazard**
- $h_P(A + t; \mathbf{z})$ - hazard due to all other causes of death: expected hazard in the general population or **background mortality**

Main estimators: non-parametric and parametric

- Non-parametric estimator for net survival: ‘gold standard’ (Perme et al., 2012)
- Parametric and semi-parametric estimators (frequentist inference):
  - Regression models on the excess hazard scale (Estève et al., 1990; Remontet et al., 2007; Charvat et al., 2016)
  - Regression models on the cumulative excess hazard scale (Lambert et al., 2009)
  - GLM formulation modelling the number of deaths (Dickman et al., 2003)

Log-likelihood

$$\sum_{i}^{t_n}\left(\delta_i.log\left(h_E(t_i;\mathbf{x}) + h_P(A+t_i;\mathbf{z})\right) - \int_0^{t_i}\left(h_E(u;\mathbf{x}) + h_P(A+u;\mathbf{z})\right)du\right)$$

$$h_E(t_i;\mathbf{x}) = h_{E_0}(t).exp\left(\sum_{j\in J}\beta_j.x_j + \sum_{k\in K} f_k(x_k) + \sum_{l\in L} g_l(t).x_l\right)$$

- $h_{E_0}(t)$ - baseline excess hazard commonly modelled using flexible functions (restricted cubic splines* or B-splines*)
- Variables in **set J** modelled with a linear effect
- Non-linear effects in **set K** modelled using flexible functions*
- Non-proportional effects in **set L** modelled including interaction between covariates and time

Added complexity: likelihood formulation with no closed-form expression

Common solution: use numerical integration rules

Choice of splines: Low-Rank Thin Plate splines

- Likelihood function remains tractable, avoiding numerical integration
- First-order polynomials - penalised splines
- Simple yet enough flexibility to capture the shapes of excess hazards

Piecewise linear log-baseline excess hazard model:

Given a partition of the follow-up time range as $0 = \tilde{t}_0 < \tilde{t}_1 < \ldots < \tilde{t}_k = \infty$

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

Implementation involves a series of transformations to the spline parameters α^*, as well as constructing a time design matrix and a penalty transformation matrix (Crainiceanu et al. 2005 and Murray et al., 2016)

Illustration: using population-based cancer data

- **Data:** All adult men diagnosed with colon cancer during 2009 in London, and followed-up until the 31th December 2015.
- **Variables available for analysis:**
  - full dates of diagnosis, last follow-up and death
  - vital status indicator (dead or alive at the end of follow-up)
  - age at diagnosis (15-99 years)
  - deprivation category (Index of Multiple Deprivation - income domain)
- **Background mortality:** Life tables for England stratified by calendar year (2009-2015), age, gender, deprivation category and region of residence

Illustration: Model set-up

Age at diagnosis (A) and deprivation category (dep) as main effects:

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

- **part 1** spline modelling the baseline log-excess hazard using 4 partitions of the observed follow-up time, and incorporating the time-dependent effect of age at diagnosis.
- **part 2** spline modelling the non-linear effect of age at diagnosis using 3 partitions (J=3) of the observed age range.
- **part 3** formulates the linear and proportional effect of deprivation.

Illustration: Bayesian estimation

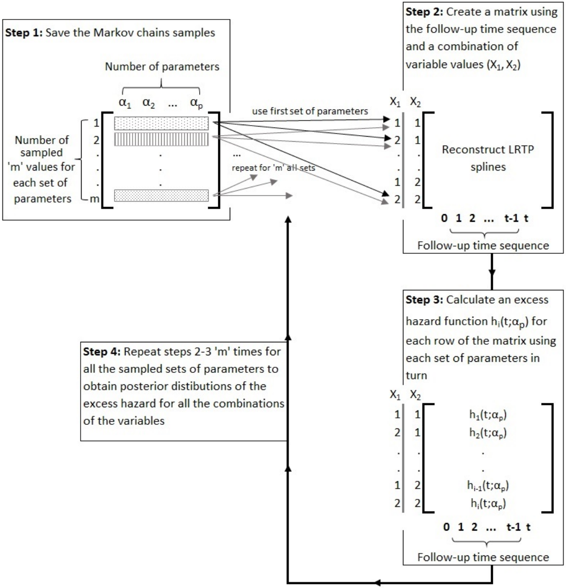
- Prior distributions used for the parameters:
  - for the baseline log-excess hazard:

$$\alpha_0 \sim N(0, 10^4),\ \alpha_1 \sim N(0, 10^4)$$

$$\alpha_k|\sigma_\alpha \overset{iid}{\sim} N(0, \sigma_\alpha^2), \text{ for k=2}, \ldots, \text{ K and } \sigma_\alpha \sim U(0.01, 100)$$

- Model fitted in `JAGS` accessed via `R2Jags`
- 30,000 MCMC samples from each posterior distribution
- Examination of trace and density plots did not indicate any convergence issues
- MCMC samples were save and posterior distributions for excess hazard and net survival were derived in a post-estimation procedure

Illustration: Calculating posterior distributions



Step 1: Save the Markov chains samples

Step 2: Create a matrix using the follow-up time sequence and a combination of variable values (X_1, X_2)

Step 3: Calculate an excess hazard function $h_i(t;\alpha_p)$ for each row of the matrix using each set of parameters in turn

Step 4: Repeat steps 2-3 'm' times for all the sampled sets of parameters to obtain posterior distibutions of the excess hazard for all the combinations of the variables

Illustration: Summarising posterior distributions

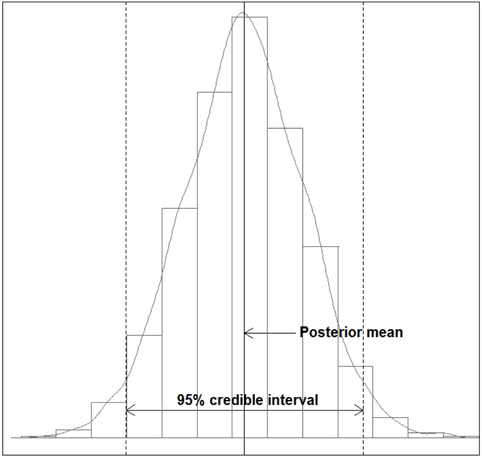
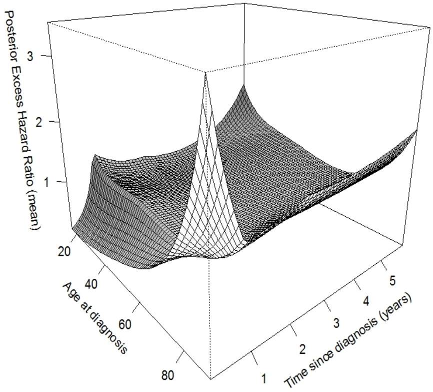


Illustration: Posterior Excess Hazard Ratios (mean)



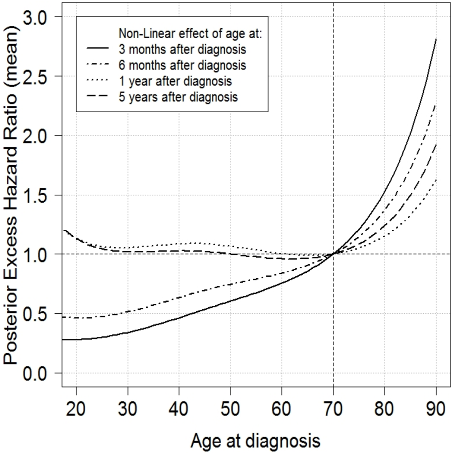
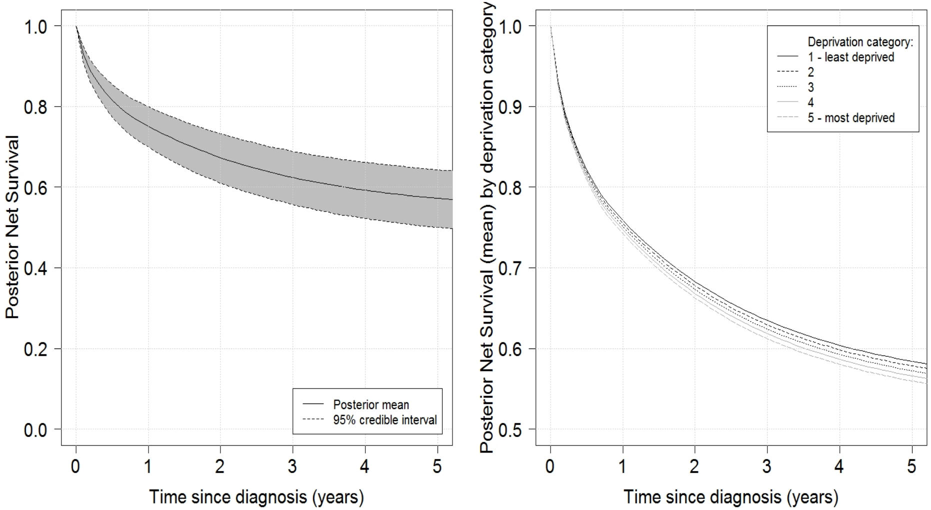
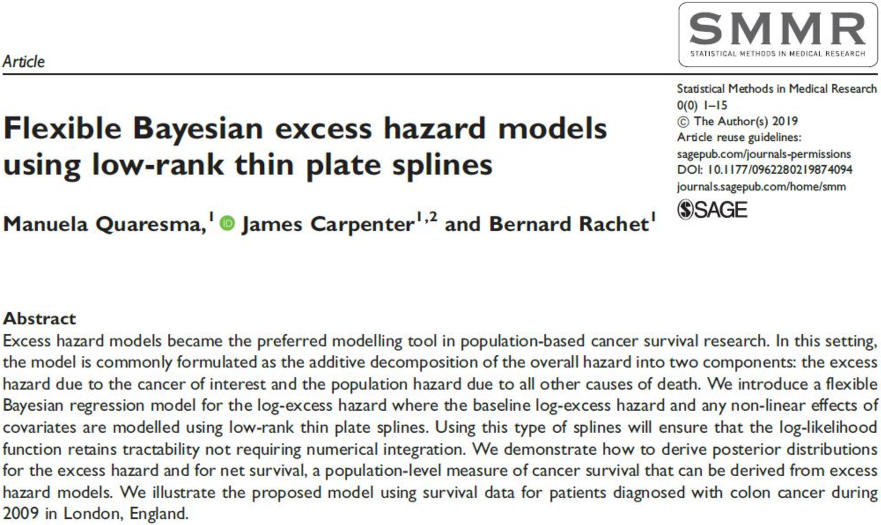


Illustration: Posterior Net survival (mean)





SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Article

Statistical Methods in Medical Research
0(0) 1–15
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280219874094
journals.sagepub.com/home/smm
SAGE

Flexible Bayesian excess hazard models using low-rank thin plate splines

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.