Bayesian variable selection and survival modeling: Assessing the most important comorbidities that impact lung and colorectal cancer survival in Spain

Francisco Javier Rubio

University College London

Danilo Alvares

Pontificia Universidad Católica de Chile

Daniel Redondo-Sanchez

Instituto de Investigación Biosanitaria

Rafael Marcos-Gragera

Institut Català d'Oncologia

María-José Sánchez

Andalusian School of Public Health

Miguel Angel Luque-Fernandez (✉ miguel-angel.luque@lshtm.ac.uk)

London School of Hygiene & Tropical Medicine

Research Article

Keywords: Bayesian variable selection, Cancer survival, Comorbidities, Conditional effects, Marginal effects

Posted Date: November 18th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1053024/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at BMC Medical Research Methodology on April 3rd, 2022. See the published version at <https://doi.org/10.1186/s12874-022-01582-0>.

RESEARCH

Bayesian variable selection and survival modeling: Assessing the most important comorbidities that impact lung and colorectal cancer survival in Spain

Francisco Javier Rubio¹, Danilo Alvares², Daniel Redondo-Sanchez^{3,4,5}, Rafael Marcos-Gragera^{4,6,7}, María-José Sánchez^{3,4,5,8} and Miguel Angel Luque-Fernandez^{3,4,5,9*}

*Correspondence:

miguel-angel.luque@ishtm.ac.uk

³ Non-communicable Disease and Cancer Epidemiology Group, Instituto de Investigación Biosanitaria de Granada (ibs.GRANADA), Granada, Spain
Full list of author information is available at the end of the article

Abstract

Cancer survival represents one of the main indicators of interest in cancer epidemiology. However, the survival of cancer patients can be affected by a number of factors, such as comorbidities, that may interact with the cancer tumour. Moreover, it is of interest to understand how different cancer sites and tumour stages are affected by different comorbidities. Identifying the comorbidities that affect cancer survival is thus of interest as it can be used to identify factors driving the survival of cancer patients. This information can also be used to identify vulnerable groups of patients with comorbidities that may lead to a worst prognosis of cancer. We address these questions and propose a principled selection and evaluation of the effect of comorbidities in the overall survival in cancer patients. In the first step, we apply a Bayesian variable selection method that can be used to identify the comorbidities that predict overall survival. In the second step, we build a general Bayesian survival model that accounts for time-varying effects. In the third step, we derive several posterior predictive measures to quantify the effect of individual comorbidities on the population overall survival. We present applications to data on lung and colorectal cancers from two Spanish population-based cancer registries. The proposed methodology is implemented with a combination of the R-packages `mombf` and `rstan`. We provide the code for reproducibility at <https://github.com/migariane/BayesVarImpComorbiCancer>.

Keywords: Bayesian variable selection; Cancer survival; Comorbidities; Conditional effects; Marginal effects

Introduction

Selecting the set of patient and tumour characteristics that better predict the survival probability of cancer patients is of primary interest in cancer epidemiology as this information can be used to inform policy makers and epidemiologists (see Michalopoulou *et al.*[1] for a discussion). Moreover, quantifying the variable importance (in our context, for predicting survival) can be used to identify the most relevant patient's characteristics that may affect their prognosis. At the population level, the information about risk factors in cancer patients can be used to stratify groups of patients at higher risk. Information about comorbidities from cancer registries is typically limited, but recent computational algorithms have allowed

the identification of comorbidities at the population level using patients hospital records.[2, 3, 4]

There are several methods that can be applied to perform variable selection (*e.g.*, comorbidities in prognostic cancer epidemiology) in the frequentist and Bayesian frameworks. Briefly, in a frequentist framework, it is common to use stepwise selection methods based on information criteria such Akaike information criterion (AIC), Bayesian information criterion (BIC), or deviance information criterion (DIC). Two limitations of this approach are their poor performance in terms of correct model selection for finite samples, and potential multiplicity problems.[5] Alternatively, penalised likelihood methods have become popular in survival analysis. These include the use of LASSO, Ridge, and Elastic nets penalties for Proportional Hazards (PH) and Accelerated Failure Time (AFT) models.[6] However, penalised methods do not allow for quantifying the uncertainty about the selected models, and its performance can be affected by high correlations between the variables. Several Bayesian variable selection methods have been proposed based on different combinations of priors and survival models (we refer the reader to Rossel and Rubio[7] for a thorough overview of these methods). These methods allow for quantifying the uncertainty about the selection models, as posterior probabilities can be assigned to each model, and to quantify the variable importance through the calculation of posterior inclusion probabilities (PIP), which can be interpreted as a measure of the importance of individual variables in explaining the response.[8]

In practice, we are interested in conducting statistical inference and drawing conclusions from the selected variables. Thus, post-selection inference is an step as important as variable selection. The natural post-selection steps are two: modelling the survival response based on the selected variables, and quantifying the effect of the selected variables on the survival probability. There is a large catalogue of survival regression models, which are typically formulated in terms of the hazard function and aim at including effects that play a role on the time-scale and the hazard-scale. We refer the reader to Rubio et al[9] for a review on hazard regression models, and for a detailed discussion on the parametric models that will be used in this paper (which we will refer to as General Hazards models). General Hazards (GH) models allow for incorporating hazard-level and time-level effects (*i.e.*, effects that play a role on the hazard and time scales) while avoiding the need for numerical integration. This class of models includes, as particular cases, the PH, the AFT, and Accelerated Hazards (AH) models. This tractability and interpretability is useful in survival data modelling as it allows the user to specify the roles of the covariates in the model. Once the survival model is fitted, it is useful to produce model summaries that help the user gain understanding on the effect of the variables on the survival probability. There exist several conditional and marginal measures that can be used to assess these effects. Briefly, conditional effect measures aim at quantifying the effect of a variable in the observed population by comparing the survival curves, at specific time points, associated to individuals with and without a characteristic of interest (*e.g.* a comorbidity). Conditional measures in the context of survival analysis include the conditional risk differences and hazard ratios,[10] the attributable risk, and the attributable survival of a particular covariate pattern.[11] In contrast, marginal measures aim at quantifying the effect of a characteristic of interest in the

entire population. These include marginal effects based on the survival function, the marginal risk differences, and the restricted mean survival time.[11] These concepts will be described in the following sections.

Based on the use of Bayesian methodology, we aim to provide the end-user with guidelines to address the research questions about: (i) how to select the most important variables (*i.e.*, comorbidities) that affect the survival of cancer patients accounting for the modeling uncertainty; and (ii) how to quantify the effect of the selected variables using conditional and marginal measures of association. We also aim to illustrate the proposed methodology using real data containing information on comorbidities and survival times of cancer patients.

The remainder of the paper is organised as follows. Section presents a discussion of the data sets that motivated our work, and that will be used in our applications. Section introduces the three steps in the proposed Bayesian setting, including variable selection, survival modelling, and the calculation of summary measures. Section presents two applications using population-based data on colorectal and lung cancer in Spain. We discuss the use and interpretation of the proposed methodology and explore the conclusions obtained by stratifying the data by grouping tumour cancer stages (using their biological and clinical differences). Since the variable containing information about tumour stage contains missing observations, we also present a sensitivity analysis of the results obtained by using complete cases *vs.* imputing the missing covariates.

Motivating Examples

The methodological framework proposed in this paper is motivated by timely and recent epidemiological questions. In this section we describe the data sets that motivates these research questions.

Worldwide, lung and colorectal cancer are currently among the three most frequent anatomical locations in terms of incidence and mortality.[12] Cancer survival indicators widely varies between countries, with the five-year age-standardised net survival in Spain during 2010-2014 being 63.2% for colon, 59.5% for rectum, and 13.5% for lung cancer.[13]

We analyse data from a population-based cohort study including patients diagnosed with colorectal cancer (CRC) and primary lung and bronchus cancer incident cases diagnosed from 1st January 2011 to 31st December 2012 in two Spanish population-based cancer registries - Girona and Granada. The diagnoses were based on codes C18-C21 for CRC and C34.0, C34.1, C34.2, C34.3, C34.8, C34.9, for lung cancer according to the International Classification of Diseases for Oncology, 3rd Edition.[14] The entry date of each patient into the cohort was defined as the date of cancer diagnosis, and their exit date was defined as the date of death or the date at 6 years after their cancer diagnosis for CRC cancer and 8-years for lung cancer, whichever occurred first.

Data on cancer stage at diagnosis (TNM staging system, 7th edition [15]), comorbidities, and sex, were obtained retrospectively from patients' medical records. The data collection followed a detailed protocol from the European High Resolution studies collaboration (TRANSCAN-HIGHCARE project within the ERA-Net [16]). The vital status was assessed using the national death registry of the Spanish Ministry of Health. Comorbidities were assessed using the codes from the International

Classification of Diseases, 10th Revision.[17] All retrospectively recorded comorbidities in the medical records were included except for those diagnosed within 6 months before cancer diagnosis to prevent including cancer related comorbidities.[2]

Methodological Framework

In this section, we describe the proposed Bayesian methodological framework. Briefly, we propose the following steps:

- 1 In step one, we select the relevant variables that predict the survival times using an AFT model coupled with two types of priors.[7] The idea behind this methodology consists of selecting variables in AFT models using priors specifically developed to improve the finite-sample performance and consistency of selection. These include, the g -Zellner prior (π_L) as well a non-local prior (π_M),[7] which help enforce parsimony. Variable selection is conducted using a formal Bayesian approach based on posterior probabilities of the different models as well as assessing the importance of the selected variables via calculating their PIPs. We perform this step using the R-package `mombf` (version 3.0.4).[18]
- 2 The second step consists of model building, which is based on a rich family of hazard regression models that contains the most common survival models (PH, AFT, and AH) as particular cases.[9] We fit this Bayesian survival model using the R-package `rstan` (version 2.21.2).[19]
- 3 In step three, we provide several conditional and marginal posterior predictive measures that allow for the quantification of the effects of the selected individual comorbidities on the population survival.

We emphasise that these steps can be conducted on the entire population, or stratified sub-populations of interest. For instance, one could think of stratifying the population into early and late tumour stages as there are biological and clinical differences between advanced and early stages of cancer[20] or sex[21]. We explore and discuss this idea further in our applications. Note that the software can be easily adapted to other data sets and the code for running the three steps in the proposed methodological approach is available at the GitHub repository: <https://github.com/migariane/BayesVarImpComorbiCancer>.

Step 1: Bayesian variable selection

Throughout, let $o_i \in \mathbb{R}_+$ be the time to event of interest, and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top \in \mathbb{R}^p$ be the corresponding vector of covariates containing all of the available patient characteristics, for individuals $i = 1, \dots, n$. Let $c_i \in \mathbb{R}_+$ denote the right-censoring times, such that one only observes the times $t_i = \min\{o_i, c_i\}$. Denote by $\delta_i = \mathbf{I}(o_i < c_i)$ the vital status observation i , and define $y_i = \min\{\log(o_i), \log(c_i)\}$ the observed log-times, $\mathbf{y} = (y_1, \dots, y_n)$, $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$, and the number of uncensored individuals $n_o = \sum_{i=1}^n \delta_i$.

The variable selection step is based on the proposed methodology in Rossell and Rubio[7], which we briefly detail below. The aim here is to select the important variables that explain survival. To do so, we introduce an inclusion indicator $\gamma =$

$(\gamma_1, \dots, \gamma_p)$, where

$$\gamma_j = \begin{cases} 0, & \text{if } \beta_j = 0, \\ 1, & \text{if } \beta_j \neq 0, \end{cases}$$

and $j = 1, \dots, p$. That is, $\gamma = (\gamma_1, \dots, \gamma_p)$ determines which covariates are included in the model. For the selection step, we adopt an AFT model. This model assumes a log-linear regression structure:

$$\log(o_i) = \mathbf{x}_i^\top \boldsymbol{\beta}_\gamma + \epsilon_i, \quad (1)$$

where ϵ_i are independent across $i = 1, \dots, n$ with mean $E(\epsilon_i) = 0$ and variance $Var(\epsilon_i) = \sigma^2$. For simplicity, we will assume that $\epsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$. Let $\mathbf{X}, \mathbf{X}_o, \mathbf{X}_c$ denote the design matrices associated to entire sample, the uncensored survival times, and the censored survival times, respectively. Throughout, we will assume that $\mathbf{X}_o$ has full column rank.

In order to obtain a log-concave likelihood function, which in turns improves the performance of optimisation methods, Rossell and Rubio[7] adopt the reparameterisation $\boldsymbol{\alpha}_\gamma = \boldsymbol{\beta}_\gamma / \sigma$, and $\tau = 1/\sigma$. The log-likelihood under this parameterisation can be written as follows

$$\ell(\boldsymbol{\alpha}_\gamma, \tau) = -\frac{n_o}{2} \log\left(\frac{2\pi}{\tau^2}\right) - \frac{1}{2} \sum_{\delta_i=1} (\tau y_i - \mathbf{x}_i^\top \boldsymbol{\alpha}_\gamma)^2 + \sum_{\delta_i=0} \log \{\Phi(\mathbf{x}_i^\top \boldsymbol{\alpha}_\gamma - \tau y_i)\}$$

We adopt the following priors for the model parameters:[7]

$$\begin{aligned} \pi_L(\boldsymbol{\alpha}_\gamma, \tau) &= \prod_{\gamma_j=1} N(\alpha_j; 0, g_L n / (x_j^\top x_j)) \pi(\tau), \\ \pi_M(\boldsymbol{\alpha}_\gamma, \tau) &= \prod_{\gamma_j=1} \frac{\alpha_j^2}{g_M} N(\alpha_j; 0, g_M) \pi(\tau), \end{aligned}$$

where $\pi(\tau) = 2\tau^{-3} \text{IG}(\tau^{-2}; a_\tau/2, b_\tau/2)$, and IG denotes the inverse gamma density, and $g_L, g_M, a_\gamma, b_\tau \in \mathbb{R}_+$ are given dispersion parameters. Following the default recommendations in [7], we use $g_M = 0.192$, $g_L = 1$, and $a_\tau = b_\tau = 1$ in our applications, which assign low probability to small effects with $e^{\beta_j} < 1.15$. We also adopt a Beta-Binomial prior on the different models: [7]

$$\pi(\gamma) = \text{BetaBin}(p_\gamma; p, a_1, b_1),$$

where $\text{BetaBin}(z; p, a, b)$ is the probability of z successes under a Beta-Binomial distribution with p trials and parameters (a, b) .

Based on this formulation, we obtain the following model posterior probabilities

$$\pi(\gamma \mid y) = \frac{p(y \mid \gamma) \pi(\gamma)}{\sum_{\gamma} p(y \mid \gamma) \pi(\gamma)} = \left(1 + \sum_{\gamma' \neq \gamma} B_{\gamma', \gamma} \frac{\pi(\gamma')}{\pi(\gamma)} \right)^{-1}, \quad (3)$$

where $\pi(\gamma)$ is the model prior probability, $B_{\gamma',\gamma} = p(y \mid \gamma')/p(y \mid \gamma)$ the Bayes factor between (γ', γ) and

$$p(y \mid \gamma) = \int p(y \mid \alpha_\gamma, \tau) \pi(\alpha_\gamma, \tau \mid \gamma) d\alpha_\gamma d\tau,$$

the integrated likelihood $p(y \mid \alpha_\gamma, \tau)$ with respect to a prior density $\pi(\alpha_\gamma, \tau \mid \gamma)$. This integrated likelihood is calculated with a Laplace approximation in Rossell and Rubio[7].

One option for model selection consists of choosing the model with highest posterior probability $\hat{\gamma} = \operatorname{argmax}_\gamma \pi(\gamma \mid y)$. However, for models with many covariates, the 2^p models are often assigned low probabilities. Thus, the posterior model probabilities are combined with marginal PIPs of the variables.[22]

$$\text{PIP}(\gamma_j = 1 \mid y) = \sum_{\gamma_j=1} \pi(\gamma \mid y),$$

which represents the sum of the posterior probabilities of the models that contain the variable of interest $\gamma_j = 1$. This quantity can be used to assess the individual variable importance, and has a nice interpretation as it has a formal connection with a probability (thus, naturally bounded on the interval $[0,1]$). It is often useful to look at the model containing those variables with PIP larger than 0.5, which can be used to build a survival model for the selected variables. This methodology is implemented in the R-package **mombf** (version 3.0.4).[18]

Step 2: Modelling using a Bayesian flexible parametric hazard regression

Once the set of important variables, say $\mathbf{z}_i \in \mathbb{R}^q$, are selected, we develop a richer survival model that allows for the inclusion of time-dependent effects, hazard-level effects, as well as a flexible parametric baseline hazard. More specifically, we consider the general hazard structure

$$h_{GH}(t \mid \mathbf{z}_i, \boldsymbol{\xi}, \tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = h_0 \left(t \exp\{\tilde{\mathbf{z}}_i^\top \tilde{\boldsymbol{\theta}}\} \mid \boldsymbol{\xi} \right) \exp\{\mathbf{z}_i^\top \boldsymbol{\theta}\}, \quad (4)$$

where $h_0(\cdot \mid \boldsymbol{\xi})$ is a parametric baseline hazard with vector parameter $\boldsymbol{\xi} \in \Xi$, $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{\tilde{q}}$ represents the regression coefficients associated to the time-dependent effects $\tilde{\mathbf{z}}_i \in \mathbb{R}^{\tilde{q}}$, $\boldsymbol{\theta} \in \mathbb{R}^q$ are the regression coefficients associated to the hazard-level effects $\mathbf{z}_i$. Typically, $\tilde{\mathbf{z}}_i \subset \mathbf{z}_i$. This hazard structure contains, as particular cases, the Proportional Hazards (PH, $\tilde{\boldsymbol{\theta}} = 0$), Accelerated Hazards (AH, $\boldsymbol{\theta} = 0$), and Accelerated Failure Time (AFT, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and $\tilde{\mathbf{z}}_i = \mathbf{z}_i$) models. The baseline hazard can be chosen to be the hazard associated to the 3-parameter Power Generalised Weibull or Generalised Gamma distributions which can capture the basic shapes of interest in practice (increasing, decreasing, unimodal, and bathtub).[23] Simpler 2-parameter distributions such as the Log-Normal, Log-Logistic, or Gamma distributions can be used as well. In our implementation, we allow for a variety of combinations of baseline hazards. Moreover, since the implementation is done in **rstan** (version 2.21.2),[19] this allows for selecting the best survival model (*i.e.*, the combination of hazard structure and parametric baseline hazard) using posterior model probabilities calculated with the R-package **bridgesampling** (version

1.1-2).[24] Since the aim of this step consists of inference on the model parameters, we adopt weakly informative priors. This requires a case-by-case analysis, but in general, we adopt half-Cauchy priors for scale and shape parameters of the baseline hazard, and normal priors with large variance for the regression coefficients. We point out that other priors could be used as well (see Alvares and Rubio[25] for a discussion), and that our implementation in **rstan** allows for easily changing this choice. In our applications, we only include comorbidities in $\mathbf{z}$ as we do not expect binary covariates to have a time-varying effect (based on discussions with clinicians). This is, we only consider hazard-level effects of the comorbidities as these are binary variables.

Step 3: Summary measures

The Bayesian model fitted in Step 2 will now be used to explore the effect of the comorbidities using several predictive posterior conditional and marginal measures of association between the vector of selected comorbidities and cancer survival. We start by calculating the conditional posterior hazard ratios (HR) of the comorbidities, which are simply obtained as the exponential of the corresponding estimates of the coefficients, based on the hazard structure (4). We now introduce two survival functions, which represent posterior predictive conditional and marginal population survival functions associated to a comorbidity of interest. Let $(\boldsymbol{\xi}^{(j)}, \tilde{\boldsymbol{\theta}}^{(j)}, \boldsymbol{\theta}^{(j)})$, $j = 1, \dots, M$, be a posterior sample of the model parameters. The predictive posterior survival function conditional on $z_{i,k} = r$ is defined by:

$$\text{CS}(t, k, r) = \frac{1}{n_r M} \sum_{j=1}^M \sum_{z_{i,k}=r} \exp \left\{ -H_{GH} \left(t \mid \mathbf{z}_i, \boldsymbol{\xi}^{(j)}, \tilde{\boldsymbol{\theta}}^{(j)}, \boldsymbol{\theta}^{(j)} \right) \right\}, \quad (5)$$

where n_r are the number of individuals with $z_{i,k} = r$, $r \in \{0, 1\}$, and H_{GH} represents the cumulative hazard function associated to (4), which does not require numerical integration.

Now, let $z_{i,k}$ be comorbidity of interest for patient i , and let $\mathbf{z}_{i,-k}$ and $\tilde{\mathbf{z}}_{i,-k}$ be the vectors of covariates after removing the covariate of interest $z_{i,k}$. The predictive posterior marginal survival function associated to assuming that comorbidities $z_{i,k} = r$, $r = 0, 1$ and $i = 1, \dots, n$, is defined as follows:

$$\text{MS}(t, k, r) = \frac{1}{nM} \sum_{j=1}^M \sum_{i=1}^n \exp \left\{ -H_{GH} \left(t \mid \mathbf{z}_{i,-k}, z_{i,k} = r, \boldsymbol{\xi}^{(j)}, \tilde{\boldsymbol{\theta}}^{(j)}, \boldsymbol{\theta}^{(j)} \right) \right\}. \quad (6)$$

These two predictive survival functions (*i.e.*, conditional and marginal survival functions) represent the main ingredient in the following measures of the effect of the r th comorbidity on the overall population survival. Note that we do not favour the use of a specific measure, but we consider that both can provide complementary information.

From (5) and (6) we can compute the following conditional or marginal measures:

- (i) The posterior predictive conditional effect (PPCE) is an effect on the conditional survival function. This is simply obtained as the difference of the

conditional survival functions at time t for the comorbidity of interest:

$$\text{PPCE}(t, k) = \text{CS}(t, k, 0) - \text{CS}(t, k, 1).$$

This function is interpreted at a population level and represents the change in survival in the group of patients having a particular comorbidity (k) compared to the group of cancer patients who do not have that comorbidity. A positive PPCE, quantifies the effect of a comorbidity in reducing survival in the exposed sub-population compared to the non-exposed sub-population.

- (ii) The posterior predictive attributable survival (PPAS) function:

$$\text{PPAS}(t, k) = \pi_r \frac{\text{CS}(t, k, 0) - \text{CS}(t, k, 1)}{\text{CS}(t, k, 0)} = \pi_r \left[1 - \frac{\text{CS}(t, k, 1)}{\text{CS}(t, k, 0)} \right],$$

where π_r is the proportion of exposed individuals in the population to the r th comorbidity. This is a time-varying weighted relative difference of the conditional survival functions, which represents the ratio of the change in survival in the exposed *vs.* non-exposed groups (for a specific comorbidity) and the conditional survival of the non-exposed group.

- (iii) The posterior predictive attributable risk (PPAR) function:

$$\text{PPAR}(t, k) = \frac{\pi_r [\text{CS}(t, k, 0) - \text{CS}(t, k, 1)]}{\pi_r [\text{CS}(t, k, 0) - \text{CS}(t, k, 1)] + 1 - \text{CS}(t, k, 0)}.$$

This is also a time-varying weighted version of the PPCE. In our context, this function can be interpreted as the portion of detrimental outcome rate attributable to a comorbidity. We refer the reader to Cox, Chu and Muñoz[26] for a more extensive discussion on the effect measures (ii) and (iii).

- (iv) The posterior predictive marginal effect (PPME) at time t is defined as

$$\text{PPME}(t, k) = \text{MS}(t, k, 0) - \text{MS}(t, k, 1).$$

This function represents the change in survival in the entire population induced from having a particular comorbidity (k). This measure takes values in $(-1, 1)$ and is interpreted as the marginal survival probability difference between having or not a particular comorbidity in the entire population.[27]

- (v) The posterior predictive restricted mean survival time (RMST)

$$\text{RMST}(t^*, k, r) = \int_0^{t^*} \text{MS}(t, k, r) dt.$$

The RMST represents the area under the marginal survival curve $\text{MS}(t, k, r)$ between time 0 and a horizon time t^* , which does not rely on the proportionality of hazards assumption. The RMST can be interpreted as the average time free from an event (dead) between time 0 and the horizon time t^* . Comparing the functions $\text{RMST}(t^*, k, r)$ for different time horizons, and $r = 1$ *vs.* $r = 0$ helps quantifying the effect of a comorbidity on survival along a time period.[28, 29]

Applications

In this section we present two applications of the proposed methodology in the contexts of assessing the interplay of comorbidities with colorectal and lung cancer survival. The data sets used in these applications were described in Section , so we focus on presenting and discussing results. In addition to the proposed three steps in our methodology, we address the problem of missing data, which mainly affect the variable containing information about tumour stage. We perform a sensitivity analysis by replicating the three steps after imputing the missing variables (stage and comorbidities for both CRC and lung cancer, and smoking status for lung cancer), assuming data were missing at random. We use the R-package `mice` (version 3.13.0) to implement a multivariate imputation via Chained Equations.[30] We impute five data sets. The imputation model included the Nelson-Aalen estimator of the cumulative hazard evaluated at the exit time, age, sex, and the vital status indicator. Variable importance results after the multiple imputation of missing stage were consistent and selected the same covariates.

CRC data analysis by stage

Among 1,061 CRC patients 59.5% were female with stages I-III and 62.2% with stage IV, respectively. The overall mean age at CRC diagnosis was about 70 years. The pattern of comorbidities was similar by CRC stage. The most common comorbidities among stages I-III CRC patients were diabetes (25.2%), chronic obstructive pulmonary disease (COPD) (18.5%), and heart failure (14.5%). Among stage IV patients, we observed diabetes (20.4%), heart failure (16.2%), and COPD (14.2%). There was 5.7% of missing stage (Supplementary Table 1).

Figure 1a shows the results from the Bayesian variable selection step (Step 1) for age, sex and ten comorbidities among CRC patients with stages I-III. The variables with a PIP higher than 0.5 were age, pulmonary disease and cerebrovascular disease. However, among CRC patients with stage IV, the higher PIP was for age, diabetes and renal disease (1.00, 0.44 and 0.35 PIP respectively, Figure 1b).

Based on the results from the Bayesian variable importance we fitted several Bayesian survival models based on both the general hazard (GH) structure (see Equation (4)) and proportional hazards (PH), which is a particular case of GH, as discussed in Section . For each of these structures, different baseline hazard functions are specified, such as Log-Normal (LN), Log-Logistic (LL), and 3-parameter Power Generalised Weibull (PGW).

For CRC patients with stages I-III, the proposed models were compared using posterior model probabilities (see Supplementary Table 2). The best model was the PH model with LL baseline hazard including age, cerebrovascular disease and COPD as covariates. Table 1 shows the posterior mean, hazard ratio (HR) and 95% credible intervals (CI), and the probability of an $HR > 1$ given the data for the LL model. CRC patients with stages I-III and cerebrovascular disease had approximately two times higher (95% CI: 1.38 – 3.02) risk of death at 6 years after CRC diagnosis compared to CRC patients without that comorbidity (Table 1). Evidence shows that these two comorbidities may delay or modify treatment alternatives among CRC patients, thus it may explain the higher risk of death and shorter CRC survival.[31] For CRC patients with stage IV, the best model was the

PH model with PGW baseline hazard including age and diabetes as covariates (see Supplementary Table 2). Table 2 shows the posterior mean, HR and 95% CI, and the probability of an $HR > 1$ given the data for the PGW PH model. CRC patients with stage IV and diabetes had approximately 57% (95% CI: 1.15 – 2.11) higher risk of death at 6 years after CRC diagnosis compared to CRC patients without that diabetes (Table 2). Diabetes appear to increase the risk for primary cancer recurrence.[1] The association between diabetes and shorter CRC survival we found among patients with stage IV is consistent with previous evidence.[32] It has been shown that diabetes can hide or modify cancer symptoms, thus delaying cancer diagnosis.[1]

Figure 2 shows the PPME (a), PPAR (b), PPAS (c), and the RMST (d), including their respective 95% CI for cerebrovascular disease in CRC with stages I-III and derived from the LL PH model. Overall the PPME and PPAR decreased while the portion of detrimental outcome rate attributable (PPAR) to cerebrovascular disease increased over time (Figures 2 a,b,c). The RMST for CRC patients with cerebrovascular disease was consistently smaller than for patients without this comorbidity (Figure 2d). Supplementary Figures 1 and 2 show the summary measures for the other (dichotomous) variables from LL PH (early stage = I-III) and PGW PH (late stage = IV) models for CRC data.

Lung data analysis by stage

Among 1,259 lung cancer patients 13.6% were female with stages I-III and 18.1% with stage IV, respectively. The overall mean age at cancer diagnosis was about 68 years old. The pattern of comorbidities was similar by lung cancer stage. The most common comorbidities among stages I-III lung cancer patients were COPD (42.5%), diabetes (21.1%), and heart failure (17.7%), and COPD (31.5%), diabetes (20.4%), and heart failure (15.7%) among stage IV. There was 4.4% of missing stage and 12.9% for smoking status (Supplementary Table 3).

Figure 3a shows the results from the Bayesian variable importance selection for age, sex, and ten comorbidities among lung cancer patients with stages I-III. Only age and liver disease showed a $PIP > 0.5$ but renal disease, smoking status and dementia showed $PIPs > 0.2$. However, among lung cancer patients with stage IV, the higher PIP was for age, gender and previous smoking status (1.00, 0.72 and 0.42 PIP respectively) (Figure 3b).

Based on the results from the Bayesian variable importance we fitted several Bayesian survival models based on both the general hazard (GH) structure (see Equation (4)) and PH, which is a particular case of GH. For each of these structures, different baseline hazard functions are specified, such as Log-Normal (LN), Log-Logistic (LL), and 3-parameter Power Generalised Weibull (PGW).

For lung patients with stages I-III, the proposed models were compared using posterior model probabilities (see Supplementary Table 4). The best model was the PH model with LL baseline hazard including age, smoking status, dementia, renal and liver disease as covariates. Table 3 shows the posterior mean, hazard ratio (HR) and 95% credible intervals (CI), and the probability of an $HR > 1$ given the data for the LL model. Lung cancer patients with stages I-III and current smoker had approximately two times higher (95% CI: 1.43 – 2.86) risk of death at 8 years

after cancer diagnosis compared to lung cancer patients without that comorbidity. Furthermore, compared to lung cancer patients without dementia, those affected for that comorbidity showed 85% increased risk of death at 8 years after cancer diagnosis (HR: 1.85, 95% CI: 1.08 – 3.05) (Table 3).

For lung patients with stage IV, the best model was the PH model with PGW baseline hazard including age, sex and smoking status as covariates (see Supplementary Table 4). Table 4 shows the posterior mean, HR and 95% CI, and the probability of an HR > 1 given the data for the PGW proportional hazard model. Lung cancer patients with stage IV and current smoker had approximately 72% (95% CI: 1.28 – 2.31) higher risk of death at 6 years after cancer diagnosis compared to non-smoker lung cancer patients (Table 4).

Figure 4 shows the PPME (a), PPAR (b), PPAS (c), and the RMST (d), including their respective 95% CI for current liver disease status in lung cancer patients with stages I-III and derived from the LL proportional hazard model. Overall the PPME and PPAR decreased while the portion of detrimental outcome rate attributable (PPAR) to liver disease increased over time (Figures 4a, b, c). The RMST for lung cancer patients with liver disease was consistently smaller than for patients without this comorbidity (Figure 4d). Among lung cancer patients, the most important comorbid conditions we identified were liver and renal diseases, and smoking. All these comorbidities conditions are related to preventive lifestyle behaviours such as smoking and drinking. Smoking contributes to over 80% of lung cancers in high-income countries.[33] Supplementary Figures 3 and 4 show the summary measures for the other (dichotomous) variables from LL PH (early stage = I-III) and PGW PH (late stage = IV) models for lung data.

Discussion

We have developed a principled three-step approach to select the comorbidities that affect the survival probability of cancer patients and to quantify their effect. We have adopted a Bayesian framework in all steps as this allows us to quantify uncertainty about the selected models and to easily obtain interval inferences about quantities of interest. We have made some choices about the specific methods used for variable selection, survival modelling, and effect measures. However, one of the strengths of the proposed approach is that one could opt for alternative methods on each step. For instance, for Step 1, a number of variable selection methods could be used instead, and we refer the reader to Rubio and Rossell[7] for an extensive literature reviews on these methods. Regarding Step 2, Bayesian survival modelling is open to the use of other hazard structures as well as semiparametric methods. For Step 3, we point out that alternative model summaries and effect measures in survival analysis to the ones presented here can be included in the analysis. We argue that our work will provide cancer epidemiologists, applied researchers and other stakeholders with the tools to conduct thorough analyses on the effect of comorbidities on cancer prognosis and survival, and to foster other developments in this area.

A common approach to assess the effect of comorbidity on cancer survival include the use of weighted scores such as the Charlson comorbidity index.[34] We argue that the approach we presented here can provide a more insightful clinical utility for

identifying the most important comorbidities and their effect on cancer survival. For instance, in our applications we show how to identify vulnerable groups of patients who have combinations of comorbidities that markedly affect the population survival probability. This information can be used to inform policy-makers who can develop targeted strategies to improve cancer survival.

We have found that different combinations of comorbidities affect the survival of cancer patients with different types of cancers and tumour stages, a seemingly novel result. For instance, we identify age, cerebrovascular and COPD as risk factors for shorter survival among stages I-III CRC patients, and age and diabetes for stage IV CRC patients. Interestingly age, smoking, dementia, renal and liver disease were associated with shorter survival among stages I-III lung cancer patients, and age, sex and smoking among stage IV lung cancer patients. This information has clinical value as well as it sheds light on the interplay of comorbidities with different cancer sites and their effect on population survival. We have compared the results obtained in Step 1 (Bayesian variable selection) with those obtained using Cox-LASSO, implemented in the R-package `glmnet` (version 4.1-2).[35] The results are similar, albeit, Cox-LASSO is sensitive to the choice of the penalty parameter: using the value `lambda.min` leads to the inclusion of more variables than those selected with `lambda.1se`. This emphasises the adequacy of having a methodology that allows for quantifying uncertainty in variable selection. With a global aging population, the prevalence of comorbidities is expected to increase.[36] We emphasise that the selected comorbidities represent those that affect the population survival. Thus, the discarded comorbidities may still play an important role at individual level, but they do not markedly reduce the survival at the population level.

We acknowledge the limited generalisability of the illustration, as it only included data from two population-based cancer registries. A natural extension of our work consists of analysing the most incident cancer sites at a national level for other countries. Information about comorbidities has only recently become available at the population level in several countries.

Declarations

Abbreviations

AFT: Accelerated Failure Time AH: Accelerated Hazard AIC: Akaike Information Criteria BIC: Bayesian Information Criteria DIC: Deviance Information Criteria CI: Credible Interval COPD: Chronic obstructive pulmonary disease CRC: Colorectal Cancer CS: Posterior predictive conditional survival function GH: General Hazard LL: Log-logistic LN: Log-normal MS: Posterior predictive marginal survival function PPAR: Posterior predictive attributable risk function PPAS: Posterior predictive attributable survival function PPCE: Posterior predictive conditional effect PGW: Power Generalized Weibull PPME: Posterior predictive marginal effect PH: Proportional Hazard PIP: Posterior Inclusion Probability RMST: Restricted mean survival time

Ethics approval and consent to participate

The officials from the Department of Health of the Andalusian Regional Government approved the study and granted access to the raw data. The study proposal (CP17/00206) was approved by the internal review board of the Andalusian School of Public Health and the ethics committee from the Department of Health of the Andalusian Regional Government (study 0072-N-18). This entire study and the research protocol for involving human data was in accordance with the guidelines of the Declaration of Helsinki. No human samples were used. All data accessed for the study were fully anonymised and the informed consent was waived by the ethics committee of the Department of Health of the Andalusian Regional Government.

Consent for publication

Not applicable

Availability of data and materials

The data that support the findings of this study are available from the Regional Government of Andalusia and the Andalusian Health Department, but restrictions apply to the availability of these data, which is often the case with cancer registry data, and so are not publicly available. The Regional Government of Andalusia and the Andalusian Health Department should be contacted to access the raw data from the present study. The code and packages used in the article are made available at a GitHub repository: <https://github.com/migariane/BayesVarImpComorbiCancer>

Competing interests

The authors declare that they have no competing interests.

Funding

Miguel Angel Luque-Fernandez is supported by a Miguel Servet I Investigator award (Grant CP17/00206) and a project grant EU-FEDER-FIS PI-18/01593 from the Instituto de Salud Carlos III, Madrid, Spain. Danilo Alvares is supported by the National Fund for Scientific and Technological Development (FONDECYT, Chile) grant number 11190018.

Authors' contributions

FJR, DA, and MALF developed the concept and design of the study. MJS and RG obtained the data. FJR and DA carried the main analyses. MALF and DRS helped with supplementary analysis and the curation of the data. FJR, MALF, DA, and DRS wrote the manuscript. All authors interpreted the data, drafted, revised the manuscript and results critically. All authors read and approved the final version of the manuscript. MALF is the guarantor of the paper.

Acknowledgments

We would like to thank all cancer patients and staff employed in the cancer registries Girona and Granada, without whom this study would not be possible.

Authors' information

MALF is the senior author

Author details

¹ Department of Statistical Science, University College London, London, UK. ² Department of Statistics, Pontificia Universidad Católica de Chile, Macul, Chile. ³ Non-communicable Disease and Cancer Epidemiology Group, Instituto de Investigación Biosanitaria de Granada (ibs.GRANADA), Granada, Spain. ⁴ Biomedical Network Research Centers of Epidemiology and Public Health (CIBERESP), Madrid, Spain. ⁵ Andalusian School of Public Health, Granada, Spain. ⁶ Epidemiology Unit and Girona Cancer Registry, Oncology Coordination Plan, Department of Health, Autonomous Government of Catalonia, Catalan Institute of Oncology, Girona, Spain. ⁷ Descriptive Epidemiology, Genetics and Cancer Prevention Group, Biomedical Research Institute (IDIBGI), Girona, Spain. ⁸ Department of Preventive Medicine and Public Health, University of Granada, Granada, Spain. ⁹ Inequalities in Cancer Outcomes Network, Department of Non-communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK.

References

1. Michalopoulou, E., Matthes, K.L., Karavasiloglou, N., Wanner, M., Limam, M., Korol, D., Held, L., Rohrmann, S.: Impact of comorbidities at diagnosis on the 10-year colorectal cancer net survival: A population-based study. *Cancer Epidemiology* **73**, 101962 (2021)
2. Maringe, C., Fowler, H., Rachet, B., Luque-Fernandez, M.A.: Reproducibility, reliability and validity of population-based administrative health data for the assessment of cancer non-related comorbidities. *PLoS ONE* **12**(3), 1–14 (2017). doi:10.1371/journal.pone.0172814
3. Luque-Fernandez, M.A., Gonçalves, K., Salamanca-Fernández, E., Redondo-Sanchez, D., Lee, S.F., Rodríguez-Barranco, M., Carmona-García, M.C., Marcos-Gragera, R., Sánchez, M.J.: Multimorbidity and short-term overall mortality among colorectal cancer patients in Spain: A population-based cohort study. *European Journal of Cancer* **129**, 4–14 (2020). doi:10.1016/j.ejca.2020.01.021
4. Luque-Fernandez, M.A., Redondo-Sanchez, D., Lee, S.F., Rodríguez-Barranco, M., Carmona-García, M.C., Marcos-Gragera, R., Sánchez, M.J.: Multimorbidity by patient and tumor factors and time-to-surgery among colorectal cancer patients in Spain: A population-based study. *Clinical Epidemiology* **12**, 31–40 (2020). doi:10.2147/CLEP.S229935
5. Maity, A.K., Basu, S., Ghosh, S.: Bayesian criterion-based variable selection. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **70**(4), 835–857 (2021). doi:10.1111/rssc.12488
6. Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N.A., Trollor, J., Brodaty, H.: A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports* **10**(1), 1–10 (2020). doi:10.1038/s41598-020-77220-w
7. Rossell, D., Rubio, F.J.: Additive Bayesian variable selection under censoring and misspecification. *arXiv preprint arXiv:1907.13563* (2019)
8. Berger, J.O., Molina, G.: Posterior model probabilities via path-based pairwise priors. *Statistica Neerlandica* **59**(1), 3–15 (2005). doi:10.1111/J.1467-9574.2005.00275.X
9. Rubio, F.J., Remontet, L., Jewell, N.P., Belot, A.: On a general structure for hazard-based regression models: An application to population-based cancer research. *Statistical Methods in Medical Research* **28**(8), 2404–2417 (2019). doi:10.1177/0962280218782293

10. Stensrud, M.J., Aalen, J.M., Aalen, O.O., Valberg, M.: Limitations of hazard ratios in clinical trials. *European Heart Journal* **40**(17), 1378–1383 (2019). doi:10.1093/EURHEARTJ/EHY770
11. Cox, C., Chu, H., Muñoz, A.: Survival attributable to an exposure. *Statistics in Medicine* **28**(26), 3276–3293 (2009). doi:10.1002/SIM.3705
12. International Agency for Research on Cancer and World Health Organization: Global cancer observatory, cancer today (2021). <https://gco.iarc.fr/today/home>
13. Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Nikšić, M., Bonaventure, A., Valkov, M., Johnson, C.J., Estève, J., Ogundbiyi, O.J., Azevedo e Silva, G., Chen, W.Q., Eser, S., Engholm, G., Stiller, C.A., Monnereau, A., Woods, R.R., Visser, O., Lim, G.H., Aitken, J., Weir, H.K., Coleman, M.P., CONCORD Working Group: Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet* **391**(10125), 1023–1075 (2018). doi:10.1016/S0140-6736(17)33326-3
14. World Health Organization: International Classification of Diseases for Oncology, Third Edition (2013)
15. International Union Against Cancer: TNM Classification of Malignant Tumours, Seventh Edition, (2009)
16. TRANSCAN-2 Objectives and Partners — Transcan-2 translational cancer research program. <https://www.transcanfp7.eu/index.php/partners/transcan-2-partners.html> accessed 01.03.2021.
<https://www.transcanfp7.eu/index.php/partners/transcan-2-partners.html>
17. World Health Organization (WHO): ICD-10: International Statistical Classification of Diseases and Related Health Problems: 10th Revision, (1990)
18. Rossell, D., Cook, J.D., Telesca, D., Roebuck, P.: mombf: Bayesian Model Selection and Averaging for Non-local and Local Priors. R-package version 3.0.4, <https://cran.r-project.org/package=mombf> (2021). R-package version 3.0.4
19. Stan Development Team: RStan: the R Interface to Stan. <http://mc-stan.org/> (2020)
20. Lambert, A.W., Pattabiraman, D.R., Weinberg, R.A.: Emerging biological principles of metastasis. *Cell* **168**(4), 670–691 (2017). doi:10.1016/J.CELL.2016.11.037
21. Rubin, J.B., Lagas, J.S., Broestl, L., Sponagel, J., Rockwell, N., Rhee, G., Rosen, S.F., Chen, S., Klein, R.S., Imoukhuede, P., Luo, J.: Sex differences in cancer mechanisms. *Biology of Sex Differences* **11**(1) (2020). doi:10.1186/S13293-020-00291-X
22. Forte, A., Garcia-Donato, G., Steel, M.: Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *International Statistical Review* **86**(2), 237–258 (2018). doi:10.1111/INSR.12249
23. Alvares, D., Lázaro, E., Gómez-Rubio, V., Armero, C.: Bayesian survival analysis with BUGS. *Statistics in Medicine* **40**(12), 2975–3020 (2021). doi:10.1002/SIM.8933
24. Gronau, Q.F., Singmann, H., Wagenmakers, E.J.: bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software* **92**(10) (2020). doi:10.18637/JSS.V092.I10
25. Alvares, D., Rubio, F.J.: A tractable Bayesian joint model for longitudinal and survival data. *Statistics in Medicine* **40**(19), 4213–4229 (2021). doi:10.1002/sim.9024
26. Cox, C., Chu, H., Schneider, M.F., Muñoz, A.: Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine* **26**(23), 4352–4374 (2007). doi:10.1002/SIM.2836
27. Syriopoulou, E., Rutherford, M.J., Lambert, P.C.: Marginal measures and causal effects using the relative survival framework. *International Journal of Epidemiology* **49**(2), 619–628 (2020). doi:10.1093/IJE/DYZ268
28. Belot, A., Ndiaye, A., Luque-Fernandez, M.A., Kipourou, D.K., Maringe, C., Rubio, F.J., Rachet, B.: 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**, 53–65 (2019). doi:10.2147/CLEP.S173523
29. Royston, P., Parmar, M.K.B.: The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in Medicine* **30**(19), 2409–2421 (2011). doi:10.1002/SIM.4274
30. van Buuren, S.: Flexible Imputation of Missing Data. Chapman and Hall/CRC, ??? (2018)
31. Mounce, L.T.A., Price, S., Valderas, J.M., Hamilton, W.: Comorbid conditions delay diagnosis of colorectal cancer: a cohort study using electronic primary care records. *British Journal of Cancer* **116**(12), 1536–1543 (2017). doi:10.1038/bjc.2017.127
32. Luque-Fernandez, M.A., Gonçalves, K., Salamanca-Fernández, E., Redondo-Sanchez, D., Lee, S.F., Rodríguez-Barranco, M., Carmona-García, M.C., Marcos-Gragera, R., Sánchez, M.-J.: Multimorbidity and short-term overall mortality among colorectal cancer patients in Spain: A population-based cohort study. *European Journal of Cancer* **129**, 4–14 (2020). doi:10.1016/J.EJCA.2020.01.021
33. Nikšić, M., Redondo-Sanchez, D., Chang, Y.-L., Rodríguez-Barranco, M., Exposito-Hernandez, J., Marcos-Gragera, R., Oliva-Poch, E., Bosch-Barrera, J., Sanchez, M.-J., Luque-Fernandez, M.A.: The role of multimorbidity in short-term mortality of lung cancer patients in Spain: a population-based cohort study. *BMC Cancer* **21**:1 21(1), 1–12 (2021). doi:10.1186/S12885-021-08801-9
34. Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C.R.: A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* **40**(5), 373–383 (1987)
35. Friedman, J.H., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33**(1) (2010). doi:10.18637/jss.v033.i01
36. Fowler, H., Belot, A., Ellis, L., Maringe, C., Luque-Fernandez, M.A., Njagi, E.N., Navani, N., Sarfati, D., Rachet, B.: Comorbidity prevalence among cancer patients: a population-based cohort study of four cancers. *BMC Cancer* **20**:1 20(1), 1–15 (2020). doi:10.1186/S12885-019-6472-9

Figures

Figure 1 (a) Posterior Inclusion Probability for colorectal cancer patients with: a) Stages I-III; b) Stage IV.

Figure 2 Posterior predictive (a) marginal effect (ME), (b) attributable risk (AR), (c) attributable survival (AS), (d) restricted mean survival time (RMST), and their respective 95% credible intervals for cerebrovascular disease in early-stage cancer patients from CRC data, using the proportional hazard model with the Log-Logistic (LL) baseline hazard specification.

Figure 3 (a) Lung - Stages I-III: Posterior Inclusion Probability (PIP); (b) Lung - Stage IV: PIP.

Figure 4 Posterior predictive (a) marginal effect (ME), (b) attributable risk (AR), (c) attributable survival (AS), (d) restricted mean survival time (RMST), and their respective 95% credible intervals for current liver disease in early-stage cancer patients from lung data, using the proportional hazard model with the Log-Logistic (LL) baseline hazard specification.

Tables

Table 1 Posterior summary for the proportional hazard model with the Log-Logistic (LL) baseline hazard specification for the CRC data considering stages I-III cancer patients in Spain, $n = 770$.

Interpretation	Parameter	Posterior Mean	HR	CI 2.5%	CI 97.5%	P(HR > 1 Data)
Age	β_1	0.733	2.081	1.749	2.464	1.000
Cerebrovascular disease	β_2	0.735	2.085	1.379	3.023	1.000
COPD	β_3	0.623	1.865	1.401	2.446	1.000
LL scale	η	29.244	–	20.436	42.349	–
LL shape	ν	0.695	–	0.616	0.777	–

Table 2 Posterior summary for the proportional hazard model with the 3-parameter Power Generalised Weibull (PGW) baseline hazard specification for the CRC data considering stage IV cancer patients in Spain, $n = 287$.

Interpretation	Parameter	Posterior Mean	HR	CI 2.5%	CI 97.5%	P(HR > 1 Data)
Age	β_1	0.522	1.685	1.439	1.973	1.000
Diabetes	β_2	0.449	1.567	1.147	2.113	1.000
PGW scale	η	3.289	–	0.679	13.317	–
PGW shape 1	ν	0.793	–	0.658	0.964	–
PGW shape 2	δ	0.931	–	0.341	1.719	–

Table 3 Posterior summary for the proportional hazard model with the Log-Logistic (LL) baseline hazard specification for the lung data considering stage I-III cancer patients in Spain, $n = 566$.

Interpretation	Parameter	Posterior Mean	HR	CI 2.5%	CI 97.5%	P(HR > 1 Data)
Age	β_1	0.439	1.551	1.382	1.751	1.000
Previous smoker	β_2	0.393	1.481	1.071	2.100	0.992
Current smoker	β_3	0.699	2.012	1.426	2.863	1.000
Dementia	β_4	0.613	1.846	1.077	3.048	0.989
Renal disease	β_5	0.512	1.669	1.122	2.413	0.994
Liver disease	β_6	0.518	1.679	1.088	2.509	0.990
LL scale	η	3.221	–	1.988	5.112	–
LL shape	ν	0.943	–	0.862	1.035	–

Table 4 Posterior summary for the proportional hazard model with the 3-parameter Power Generalised Weibull (PGW) baseline hazard specification for the lung data considering stage IV cancer patients in Spain, $n = 693$.

Interpretation	Parameter	Posterior Mean	HR	CI 2.5%	CI 97.5%	P(HR > 1 Data)
Age	β_1	0.309	1.362	1.243	1.501	1.000
Female vs. male	β_2	0.170	1.185	0.928	1.543	0.910
Previous smoker	β_3	0.428	1.534	1.132	2.054	0.999
Current smoker	β_4	0.542	1.719	1.284	2.309	1.000
PGW scale	η	0.336	–	0.198	0.581	–
PGW shape 1	ν	1.150	–	0.993	1.324	–
PGW shape 2	δ	2.245	–	1.550	3.049	–

Figures

Figure 1

(a) Posterior Inclusion Probability for colorectal cancer patients with: a) Stages I-III; b) Stage IV.

Figure 2

Posterior predictive (a) marginal effect (ME), (b) attributable risk (AR), (c) attributable survival (AS), (d) restricted mean survival time (RMST), and their respective 95% credible intervals for cerebrovascular disease in early-stage cancer patients from CRC data, using the proportional hazard model with the Log-Logistic (LL) baseline hazard specication.

Figure 3

(a) Lung - Stages I-III: Posterior Inclusion Probability (PIP); (b) Lung - Stage IV: PIP.

Figure 4

Posterior predictive (a) marginal effect (ME), (b) attributable risk (AR), (c) attributable survival (AS), (d) restricted mean survival time (RMST), and their respective 95% credible intervals for current liver disease in early-stage cancer patients from lung data, using the proportional hazard model with the Log-Logistic (LL) baseline hazard specification.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplFiles.pdf](#)