

# Bayesian adaptive B-spline estimation in proportional hazards frailty models

Emmanuel Sharef

*Cornell University  
Ithaca, NY 14853*

*Department of Operations Research & Information Engineering  
e-mail: [ess28@cornell.edu](mailto:ess28@cornell.edu)*

Robert L. Strawderman

*Cornell University  
Ithaca, NY 14853*

*Department of Biological Statistics and Computational Biology  
e-mail: [rls54@cornell.edu](mailto:rls54@cornell.edu)*

David Ruppert

*Cornell University  
Ithaca, NY 14853*

*Department of Operations Research & Information Engineering  
e-mail: [dr24@cornell.edu](mailto:dr24@cornell.edu)*

Mark Cowen

*St. Joseph Mercy Hospital  
Ann Arbor MI 48106*

*The Quality Institute and Department of Medicine  
e-mail: [cowenm@trinity-health.org](mailto:cowenm@trinity-health.org)*

Lakshmi Halasyamani

*St. Joseph Mercy Hospital  
Ann Arbor MI 48106*

*Department of Medicine  
e-mail: [halasyal@trinity-health.org](mailto:halasyal@trinity-health.org)*

**Abstract:** Frailty models derived from the proportional hazards regression model are frequently used to analyze clustered right-censored survival data. We propose a semiparametric Bayesian methodology for this purpose, modeling both the unknown baseline hazard and density of the random effects using mixtures of B-splines. The posterior distributions for all regression coefficients and spline parameters are obtained using Markov Chain Monte Carlo (MCMC). The methodology permits the use of weighted mixtures of parametric and nonparametric components in modeling the hazard function and frailty distribution; in addition, the spline knots may also be selected adaptively using reversible-jump MCMC. Simulations indicate that the method produces smooth and accurate posterior hazard and frailty density estimates. The Bayesian approach not only produces point estimators that outperform existing approaches in certain circumstances, but also offers a wealth of information about the parameters of interest in the form of MCMC samples from the joint posterior probability distribution. We illustrate the adaptability of the method with data from a study of congestive heart failure.

**Keywords and phrases:** Survival analysis, knot selection, hazard regression, frailty distribution, random effect density, reversible-jump MCMC, heart failure, re-hospitalization.

Received March 2010.

## Contents

1	Introduction . . . . .	607
2	Model structure . . . . .	610
2.1	The proportional hazards frailty model . . . . .	611
2.2	Spline specifications . . . . .	612
2.3	Bayesian modeling . . . . .	612
2.3.1	Shrinking towards parametric models . . . . .	614
2.3.2	Adaptive knot selection . . . . .	616
3	Sampling from the posterior distribution . . . . .	617
3.1	Initialization . . . . .	617
3.2	Metropolis-Hastings MCMC steps . . . . .	619
3.3	Reversible-jump MCMC for adaptive knot selection . . . . .	619
3.3.1	Knot position change (move step) . . . . .	620
3.3.2	Knot addition (birth step) . . . . .	620
3.3.3	Knot deletion (death step) . . . . .	621
4	Simulation studies . . . . .	621
4.1	Curve fitting performance . . . . .	622
4.2	Parameter estimation performance . . . . .	623
5	Example . . . . .	629
5.1	Congestive heart failure data description . . . . .	629
5.2	Model selection and fitting . . . . .	630
5.3	Choice of model specification and priors . . . . .	633
6	Discussion . . . . .	637
7	Acknowledgements . . . . .	638
A	Choice of penalty functions . . . . .	638
A.1	Gaussian penalty . . . . .	638
A.2	Penalty on second differences . . . . .	638
A.3	Penalty on the second derivative . . . . .	638
B	Choice of parametric components . . . . .	639
	Supplementary Material . . . . .	639
	References . . . . .	640

## 1. Introduction

The effects of clustering in survival data are commonly addressed using a generalization of the Cox proportional hazards regression model (Cox, 1972) known as a proportional hazards frailty model or, for the purposes of this paper, a frailty model. Specifically, dependence among cluster members is assumed to

arise through an unobserved random effect, shared by all members of a cluster, that acts multiplicatively on the hazard function (Clayton and Cuzick, 1985). The corresponding marginal hazard regression model no longer exhibits a proportional hazards structure; as a result, it is no longer possible to make direct use of the partial likelihood function (Cox, 1975) in obtaining estimates of the regression parameters. However, under the assumption that the distribution of the random effect is a member of a specified parametric family of distributions (e.g., a gamma distribution) and depending on the model specification for the baseline hazard function, one may fit either a fully parametric or semiparametric model to the data (e.g., Andersen et al., 1993, Ch. IX), permitting likelihood-based estimation of all model parameters.

Extending and building upon these frequentist maximum-likelihood methods, Bayesian approaches to the frailty model have also emerged. For example, Clayton (1991) and Aslanidou, Dey and Sinha (1998) propose modeling the baseline hazard function in a gamma frailty model with processes that respectively have independent and correlated increments; see Ibrahim, Chen and Sinha (2001) and Müller and Quintana (2004) for a review. The resulting joint posterior distribution, though generally more computationally expensive to obtain than its (asymptotic) frequentist counterpart, is often better able to accommodate unusual data features and contains a wealth of information in the form of marginal and conditional posterior distributions for all parameters of interest. However, the available methods often fail to be semiparametric involving, for example, piecewise constant baseline hazard specifications that depend on some discretization of time.

Whether inference is frequentist or Bayesian, the common practice of specifying a parametric family of random effect distributions is a frequently cited disadvantage of the frailty model. For example, the results of Kosorok, Lee and Fine (2004, Prop. 6) show that misspecification of the frailty density can result in biased estimation of regression coefficients, with the direction of this bias being unpredictable. To counteract drawbacks related to frailty density misspecification, Naskar (2008) proposes to model the baseline hazard function nonparametrically and specifies the distribution of the frailty as a Dirichlet process. In order to carry out frequentist estimation in this model, Naskar (2008) employs a finite dimensional approximation to this process that amounts to the specification of a random discrete probability measure in which the number of point masses is (where feasible) set equal to the number of clusters. Earlier, Walker and Mallick (1997) proposed a related Bayesian methodology in which the baseline hazard function is specified similarly to Aslanidou, Dey and Sinha (1998) and the natural log of the frailty is assumed to follow an unspecified probability distribution  $F$  with a Pólya tree prior distribution. In theory, the Pólya tree specification corresponds to a continuous prior distribution; in practice, its use involves the specification of a finite binary partition of the support of the frailty distribution. A poorly chosen partition has the potential to obscure important features of the frailty distribution, or even introduce spurious ones.

In addition to parametric specification of the frailty distribution, there are some potential disadvantages to using a semiparametric model specification for

the hazard function in the case of the frailty model. For example, Barker and Henderson (2005) conjecture that the use of a semiparametric model leads to systematic underestimation of the frailty variance because the survival time data primarily influence the estimation of model parameters through their rank ordering, rather than through their actual magnitudes. In a frequentist context, Barker and Henderson (2005) also provide some support for this conjecture by demonstrating the use of local polynomial smoothing methods can reduce the bias in the estimation of the frailty variance. It is not known to the authors whether an analogous drawback holds in the case of Bayesian approaches to semiparametric frailty models.

Collectively, the observations made above suggest consideration of methodology that allows for highly flexible but smooth specifications of both the frailty density and baseline hazard function. In his Ph.D. thesis, Komárek (2006) presents a Bayesian nonparametric approach to clustered survival, based on the accelerated failure time model formulation of Pan (2001) and the random effect density estimation methodology of Ghidry, Lesaffre and Eilers (2004). Both the random effects density and the event time distribution are modeled as smooth mixtures of G-splines, with computations facilitated by the use of the Gibbs sampler. The resulting error and frailty distributions are smooth, allowing for easy visualization and interpretation. In this paper, we propose a related model that remains within the proportional hazards framework (i.e., conditionally upon the frailty) while implementing the desirable features of the formulation in Komárek and Lesaffre (2008). Specifically, we propose to model the baseline hazard as a penalized mixture of B-splines and the frailty density as a penalized mixture of normalized B-splines. Regarding the respective penalties as corresponding to particular prior specifications and imposing prior distributions on various other model parameters, we consider a Bayesian approach to estimation using Markov Chain Monte Carlo (MCMC). Our model formulation and the deconvolution approach bear some resemblance to those of Staudenmayer, Ruppert and Buonaccorsi (2008) and Ruppert, Nettleton and Hwang (2007). The resulting posterior estimates of the baseline hazard and frailty density are smooth and accurate, and can correctly identify unusual frailty densities and baseline hazard forms given sufficiently large samples.

Additional model flexibility is achieved in two ways. First, inspired by the work of Hjort and Glad (1995) and Hjort and Jones (1996), we permit the inclusion of parametric components that can incorporate prior knowledge about the form of the baseline hazard function and frailty density; as shown later, the resulting spline component can be interpreted as an (additive) smooth but local departure from the specified parametric model. Second, we allow the number and position of knots for the hazard and frailty B-spline bases to be chosen in a data-dependent manner, facilitated here using a reversible jump MCMC procedure. These extensions respectively improve the performance of the proposed methods in situations when the hazard or frailty density can be well-modeled by standard parametric forms or are distinctly non-smooth in nature.

The proposed adaptive knot selection procedure is related to the work of Denison, Mallick and Smith (1998) and Biller (2000). Such “free-knot” spline

methods are popular in Bayesian curve-fitting and nonlinear regression (e.g. [Smith and Kohn, 1996](#); [DiMatteo, Genovese and Kass, 2001](#); [Lindstrom, 2002](#)). In the survival analysis setting, [Mallick, Denison and Smith \(1999\)](#) have suggested a Bayesian MARS approach building on the methods of [Kooperberg, Stone and Truong \(1995\)](#) and [LeBlanc and Crowley \(1999\)](#) for nonparametrically modeling regression coefficient effects, but found that the associated computational complexity made it impractical. Retaining a semiparametric effects structure and modeling the baseline hazard and frailty density nonparametrically allows great model flexibility while remaining practical for use with large data sets.

The remainder of this paper is organized as described below. In [Section 2](#), we introduce the necessary notation, develop the basic model structure and the aforementioned extensions, and obtain the log-posterior parameter density under an appropriate noninformative right censoring scheme. We summarize the posterior sampling algorithm and key implementation details in [Section 3](#). [Section 4](#) contains illustrative simulation results, and discusses the relative merits of our approach and existing ones. In [Section 5](#), we use our methodology to analyze data from a study of congestive heart failure patients, to demonstrate the flexibility and utility of the approach in a practical setting. We conclude with a brief discussion in [Section 6](#). Appendices A and B respectively contain examples of important penalties and useful parameterizations of parametric hazard functions and frailty distributions. The accompanying supplementary materials ([Sharef et al., 2010](#)) to this paper contain an expanded version of [Section 3](#), including the various conditional posteriors used in our Gibbs sampling algorithm and adaptive knot selection stages. The methodology described herein has been implemented in the R package `splinesurv`, available from CRAN.

REMARK: We adopt a Bayesian philosophy in the development of estimation and inference procedures in this paper. However, we also take the pragmatic view that the proposed methods provide a valuable tool for frequentist-based inference. For example, in the simulations of [Section 4](#), we regard the target of estimation as being the combination of a fixed but unknown set of regression parameters, a baseline hazard function, and a frailty density and assess the frequentist performance of several relevant posterior summaries.

## 2. Model structure

The basic proportional hazards frailty model, described in [Section 2.1](#), utilizes B-spline formulations for both the baseline hazard and frailty probability density function, along with prior specifications that may be used to encourage smoothness. In [Section 2.3.1](#), this model is extended by considering convex combinations of the aforementioned B-splines and (as appropriate) a parametric hazard or density function. The resulting model, which continues to retain the form of a proportional hazards frailty model, permits (but does not force) shrinkage towards a specific parametric hazard function or frailty distribution as well as allows one to incorporate specific prior knowledge. We describe our methods

for adaptive selection of the number and placement of the B-spline knots in Section 2.3.2. Each of the baseline hazard and frailty density curves can then be specified as some combination of a basic model and its optional extensions, indexed by a variety of priors and options. The result is a flexible family of models that allows prior knowledge about the form and smoothness of either curve to be incorporated into the model fit to the desired extent.

### 2.1. The proportional hazards frailty model

We assume that the observed data structure includes information on a failure time outcome and covariate information for  $m$  independent clusters of  $J_i \geq 1$  subjects,  $i = 1 \dots m$ . In the absence of any process that censors the observation of failure times, we suppose that each subject  $(i, j)$ ,  $i = 1 \dots m$ ,  $j = 1 \dots J_i$  experiences a failure (i.e., event) at time  $X_{ij}$ , and may have available  $p$ -dimensional fixed covariates  $\mathbf{Z}_{ij}$ . We allow for the possibility of correlation between subjects within the same cluster by introducing a set of  $m$  positive cluster-level random effects, or frailties,  $U_i$ ,  $i = 1 \dots m$ . The frailties  $U_1 \dots U_m$  are assumed to be independent and identically distributed with some density  $f(\cdot)$  that has mean one; in addition, for  $i = 1 \dots m$ , subjects within the  $i^{\text{th}}$  cluster are assumed to be conditionally independent of each other given  $U_i$ . Given a baseline hazard function  $\lambda_0(\cdot)$ , a  $p$ -dimensional vector of regression coefficients  $\boldsymbol{\beta}$ , the frailties  $\mathbf{U} = (U_1 \dots U_m)^T$  and covariate information  $\mathbf{Z} = \{\mathbf{Z}_{ij}, j = 1 \dots J_i; i = 1 \dots m\}$ , we further assume that the distribution of  $X_{ij}$ ,  $j = 1 \dots J_i$ , can be specified via the proportional hazards model

$$\lambda_{ij}(t|\mathbf{U}, \mathbf{Z}) = U_i \lambda_0(t) e^{\mathbf{Z}_{ij}^T \boldsymbol{\beta}}. \quad (2.1)$$

Evidently, if the variance of the frailty probability density  $f(\cdot)$  is zero, then  $P\{U_1 = U_2 = \dots = U_m = 1\} = 1$  and all subjects are mutually independent; otherwise, members of the same cluster are marginally dependent.

In typical applications, the observation of failure times on one or more subjects may be censored. To allow for the possibility of right-censoring, we assume that the follow-up data observed on subject  $(i, j)$  takes the form  $T_{ij} = \min\{X_{ij}, C_{ij}\}$  and  $\delta_{ij} = I(X_{ij} \leq C_{ij})$ , where  $C_{ij}$ ,  $j = 1 \dots J_i$ ;  $i = 1 \dots m$  are potential censoring times. Censoring is assumed to be independent and noninformative in the same sense as Nielsen et al. (1992) (i.e., given  $\lambda_0(\cdot)$  and  $\boldsymbol{\beta}$ ). Let  $\mathbf{T}$  and  $\boldsymbol{\delta}$  denote the vector of failure times and censoring indicators corresponding to the covariates  $\mathbf{Z}$  and frailties  $\mathbf{U}$  defined above. Under the indicated assumptions and conditionally upon  $\mathbf{Z}$ , the “full data” likelihood function for  $(\boldsymbol{\beta}, \lambda_0)$  may then be written

$$\mathcal{L}(\boldsymbol{\beta}, \lambda_0 | \mathbf{T}, \boldsymbol{\delta}, \mathbf{U}, \mathbf{Z}) = \prod_{i=1}^m \left( f(U_i) \times \prod_{j=1}^{J_i} \frac{(U_i \lambda_0(T_{ij}) e^{\mathbf{Z}_{ij}^T \boldsymbol{\beta}})^{\delta_{ij}}}{\exp(U_i \Lambda_0(T_{ij}) e^{\mathbf{Z}_{ij}^T \boldsymbol{\beta}})} \right), \quad (2.2)$$

where  $\Lambda_0(t) = \int_0^t \lambda(s) ds$  (e.g. Andersen et al., 1993, Ch. IX).

## 2.2. Spline specifications

As indicated earlier, we intend to use splines for modeling  $\lambda_0(\cdot)$  and  $f(\cdot)$ . Initially, we propose to model  $\lambda_0(\cdot)$  as a non-negative linear combination of  $K_\lambda$  B-spline basis functions  $B_{\lambda k}(\cdot)$ , parametrized for example as per [de Boor \(2001\)](#). The splines are of order  $Q_\lambda$ , defined on  $N_\lambda = K_\lambda - Q_\lambda$  interior knots  $\xi_\lambda$  distributed over the range of the event times. The splines are indexed by parameters  $\theta_\lambda$ , with the weight associated with each spline basis function being given by  $w_{\lambda k} = e^{\theta_{\lambda k}}$ . That is, for  $t \geq 0$ , the baseline hazard and cumulative hazard functions are assumed to have the representations

$$\lambda_0(t|\theta_\lambda) = \sum_{k=1}^{K_\lambda} B_{\lambda k}(t)w_{\lambda k}, \quad \Lambda_0(t|\theta_\lambda) = \sum_{k=1}^{K_\lambda} w_{\lambda k} \int_0^t B_{\lambda k}(s)ds. \quad (2.3)$$

Similarly, the frailty density  $f(\cdot)$  is initially modeled as a convex combination of normalized B-spline basis functions  $\tilde{B}_{uk}(\cdot)$  of order  $Q_u$  defined on knots  $\xi_u$  over a sufficiently large range, where

$$\tilde{B}_{uk}(x) = B_{uk}(x) \cdot \left( \int_{-\infty}^{\infty} B_{uk}(s)ds \right)^{-1} \text{ for } x \geq 0.$$

The splines are indexed by parameters  $\theta_u$ , and both the B-splines and weights are normalized to ensure that the density integrates to 1. That is, given  $\theta_u$ , we assume the frailties  $U_i$  are independent and identically distributed with probability density function  $f(\cdot|\theta_u)$ , where

$$f(x|\theta_u) = \sum_{k=1}^{K_u} \tilde{B}_{uk}(x)w_{uk}, \quad w_{uk} = \frac{\exp(\theta_{uk})}{\sum_{\ell=1}^{K_u} \exp(\theta_{u\ell})} \quad (2.4)$$

for  $x \geq 0$  and is zero otherwise. In order to ensure that (2.4) has mean one, it is further assumed that  $\theta_u$  satisfies the additional constraint

$$\sum_{k=1}^{K_u} e^{\theta_{uk}} = \int_{-\infty}^{\infty} x \left[ \sum_{k=1}^{K_u} e^{\theta_{uk}} \tilde{B}_{uk}(x) \right] dx.$$

In a frequentist approach to parameter estimation, one would now maximize the marginal likelihood function that is obtained from (2.2) by integrating out  $\mathbf{U}$ . In practice, the EM algorithm or some related variant is typically used for this purpose (e.g. [Andersen et al., 1993](#), Ch. IX).

## 2.3. Bayesian modeling

The Bayesian approach to estimation for the frailty model treats  $\mathbf{U}$  as an additional unknown set of parameters; upon introducing prior distributions for all unknown model parameters, inferences for specific parameters (including  $\mathbf{U}$ )

may be derived from the appropriate marginal posterior distribution. In the present context, a Bayesian approach requires one to specify the prior distribution for the regression parameters  $\beta$  and spline parameters  $\theta = (\theta_\lambda, \theta_u)$ .

We assume that  $\beta$ ,  $\theta_\lambda$ , and  $\theta_u$  are independent. The prior distribution for  $\beta$  is assumed to be multivariate Gaussian (e.g. Staudenmayer, Ruppert and Buonaccorsi, 2008); the prior distributions for  $\theta_\lambda$  and  $\theta_u$  also have a Gaussian structure, but may incorporate a specific penalty to induce smoothness in the B-spline coefficients and avoid overfitting. Specifically, denoting the penalty functions as  $p_\lambda(\theta_\lambda)$  and  $p_u(\theta_u)$ , the corresponding priors may be written

$$\pi(\beta|\sigma_\beta^2) = (2\pi\sigma_\beta^2)^{-\frac{p}{2}} \exp\left(-\frac{1}{2\sigma_\beta^2}\beta^T\beta\right), \quad (2.5)$$

$$\pi(\theta_\lambda|\sigma_\lambda^2) \propto (2\pi\sigma_\lambda^2)^{-\frac{\kappa_\lambda}{2}} \exp\left(-\frac{1}{2\sigma_\lambda^2}p_\lambda(\theta_\lambda)\right), \quad (2.6)$$

$$\pi(\theta_u|\sigma_u^2) \propto (2\pi\sigma_u^2)^{-\frac{\kappa_u}{2}} \exp\left(-\frac{1}{2\sigma_u^2}p_u(\theta_u)\right). \quad (2.7)$$

The penalty functions  $p_u(\cdot)$  and  $p_\lambda(\cdot)$  may be chosen to follow a simple quadratic form, as in (2.5), or to penalize second differences in the parameters  $\theta_u, \theta_\lambda$ , the integrated squared second derivative of the spline, or chosen to impose some other form of smoothness criteria. Examples of possible penalty functions are presented in Appendix A. Lastly, we assume inverse-gamma priors for the error variance parameters  $\sigma_\beta^2, \sigma_\lambda^2$ , and  $\sigma_u^2$ ; for example, we have

$$\pi(\sigma_\beta^2|\alpha_\beta) \propto (\sigma_\beta^2)^{-(\alpha_{\beta 1}+1)} \exp\left(-\frac{\alpha_{\beta 2}}{\sigma_\beta^2}\right) \quad (2.8)$$

for a fixed hyperparameter  $\alpha_\beta = (\alpha_{\beta 1}, \alpha_{\beta 2})^T$ , and with analogous definitions for the priors on  $\sigma_\lambda^2$  and  $\sigma_u^2$ . These prior distributions may be quite diffuse.

REMARK: The prior distributions (2.6) and (2.7) corresponding to specific choices of the penalty functions  $p_u(\cdot)$  and  $p_\lambda(\cdot)$  may be improper. Improper priors, while not uniformly accepted among Bayesians, are commonly used in statistics. There are many reasons for this; Ghosh, Delampady and Samanta (2006, Ch. 5) provide a nice discussion. In our setting, the benefits of using an improper prior for either (2.6) or (2.7) stem mainly from the ability to incorporate smoothness and related qualitative criteria in a natural way. Suitably noninformative choices, provided one is led to a proper posterior distribution, may also lead to results that correspond more closely to those that would be obtained using a related frequentist procedure (e.g., penalized likelihood). However, because the use of improper priors also can create difficulties with the behavior and convergence of MCMC schemes (Robert and Casella, 2004), care is required in the selection of these penalty functions.

Under the proposed model specification, and with  $\sigma$  denoting the set of prior variance parameters  $(\sigma_\beta^2, \sigma_\lambda^2, \sigma_u^2)$ , the log-posterior density (up to a constant) for the parameters of interest may be written



$$\begin{aligned} \ell(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\sigma} | \mathbf{T}, \boldsymbol{\delta}, \mathbf{Z}) \\ = \sum_{i,j} \delta_{ij} [\log(U_i) + \log \lambda_0(T_{ij} | \boldsymbol{\theta}_\lambda)] \end{aligned} \quad (2.9)$$

$$- \sum_{i,j} \delta_{ij} U_i \Lambda_0(T_{ij} | \boldsymbol{\theta}_\lambda) e^{\mathbf{Z}_{ij}^T \boldsymbol{\beta}} + \sum_i \log f(U_i | \boldsymbol{\theta}_u) \quad (2.10)$$

$$- \left( \frac{p}{2} \log \sigma_\beta^2 + \frac{\boldsymbol{\beta}^T \boldsymbol{\beta}}{2\sigma_\beta^2} \right) - \sum_{d \in \{\lambda, u\}} \left[ \frac{K_d}{2} \log \sigma_d^2 + \frac{p_d(\boldsymbol{\theta}_d)}{2\sigma_d^2} \right] \quad (2.11)$$

$$- \sum_{d \in \{\beta, \lambda, u\}} \left[ (\alpha_{d1} + 1) \log \sigma_d^2 + \frac{\alpha_{d2}}{\sigma_d^2} \right], \quad (2.12)$$

with the terms in (2.9) and (2.10) corresponding to the likelihood (2.2) with  $f(\cdot)$  replaced by (2.4), the terms in (2.11) corresponding to the prior distributions (2.5)–(2.7), and the terms (2.12) corresponding to the inverse gamma prior distributions for  $\sigma_\beta^2$ ,  $\sigma_\lambda^2$ , and  $\sigma_u^2$ . (e.g., see (2.8)).

### 2.3.1. Shrinking towards parametric models

One advantage of the Bayesian approach is the possibility of supplementing the observed data with prior knowledge. In fact, the selection of a specific set of prior distributions for use in (2.9)–(2.12) implicitly defines a target towards which the posterior means of (2.3) and (2.4) are respectively to be shrunk. For example, the use of a mean zero multivariate Gaussian prior distribution shrinks the baseline hazard function (2.3) towards a constant function. Such targets have little influence on the model fit in cases where data contain substantial information about the shape of the underlying curve. However, in the absence of strong guidance from the data, these implicit targets can play a more significant role and it may be advantageous to define a more explicit target for shrinkage. Unfortunately, it is not easy to define an informative prior on the spline parameters  $\boldsymbol{\theta}$  that induces shrinkage towards a specified parametric target.

We therefore propose a modification to the spline models of Section 2.2 that permits the inclusion of a parametrically specified basis function. Specifically, we propose to replace the baseline hazard (2.3) with

$$\begin{aligned} \lambda_0(t | \boldsymbol{\theta}_\lambda, \boldsymbol{\eta}_\lambda, \phi_\lambda) &= \phi_\lambda \sum_{k=1}^{K_\lambda} B_{\lambda k}(t) w_{\lambda k} + (1 - \phi_\lambda) \lambda_{0p}(t | \boldsymbol{\eta}_\lambda) \\ \Lambda_0(t | \boldsymbol{\theta}_\lambda, \boldsymbol{\eta}_\lambda, \phi_\lambda) &= \phi_\lambda \sum_{k=1}^{K_\lambda} w_{\lambda k} \int_0^t B_{\lambda k}(s) ds + (1 - \phi_\lambda) \Lambda_{0p}(t | \boldsymbol{\eta}_\lambda), \end{aligned} \quad (2.13)$$

where  $\phi_\lambda \in [0, 1]$  and  $\lambda_{0p}(\cdot | \boldsymbol{\eta}_\lambda)$  and  $\Lambda_{0p}(\cdot | \boldsymbol{\eta}_\lambda)$  respectively denote the hazard and cumulative hazard functions for a specified parametric family of probability distributions (e.g., Weibull or lognormal). Similarly, we propose to replace the

frailty density in (2.4) with

$$f(x|\boldsymbol{\theta}_u, \boldsymbol{\eta}_u, \phi_u) = \phi_u \sum_{k=1}^{K_u} \tilde{B}_{uk}(x) w_{uk} + (1 - \phi_u) f_p(x|\boldsymbol{\eta}_u), \quad (2.14)$$

where  $\phi_u \in [0, 1]$  and  $f_p(\cdot|\boldsymbol{\eta}_u)$  denotes the density function for a specified parametric family of probability distributions (e.g., gamma or lognormal).

The baseline hazard function in (2.13) may be suggestively rewritten as

$$\lambda_0(t|\boldsymbol{\theta}_\lambda, \boldsymbol{\eta}_\lambda, \phi_\lambda) = \lambda_{0p}(t|\boldsymbol{\eta}_\lambda) + \phi_\lambda \left[ \sum_{k=1}^{K_\lambda} B_{\lambda k}(t) w_{\lambda k} - \lambda_{0p}(t|\boldsymbol{\eta}_\lambda) \right],$$

with analogous representations available for the cumulative hazard and frailty density. Consequently, we may view the proposed model extension as being equivalent to using a parametric baseline hazard, allowing for possibility of a smooth but local deviation. The prior on the weight  $\phi_\lambda$  can be used to specify the degree of confidence in the parametric component. For example, a prior favoring small values of  $\phi_\lambda$  ensures that the fit will shrink more towards the parametric specification in data-poor circumstances. The use of Beta priors on the weights  $\boldsymbol{\phi} = (\phi_\lambda, \phi_u)$  with fixed hyperparameters  $\boldsymbol{\alpha}_{\phi_\lambda}, \boldsymbol{\alpha}_{\phi_u}$  provides a simple but sensible choice. For example, a nonuniform prior such as Beta(1, 2) (triangular) ensures that the weight of the nonparametric component shrinks if it does not capture much information beyond that captured by the parametric portion. Alternatively, a Beta(2, 1) prior may be used to place additional weight on the nonparametric component if an unusual structure is suspected. Fixing  $\boldsymbol{\phi} = (1, 1)$  reduces the model to that considered in Section 2.1, whereas  $\boldsymbol{\phi} = (0, 0)$  leads to a purely parametric Bayesian survival model.

Priors for the parametric terms  $\boldsymbol{\eta}_\lambda$  and  $\boldsymbol{\eta}_u$  necessarily depend on the desired parametric form. Denote the priors by  $\pi_\lambda(\boldsymbol{\eta}_\lambda|\boldsymbol{\tau}_{\eta_\lambda})$  and  $\pi_u(\boldsymbol{\eta}_u|\boldsymbol{\tau}_{\eta_u})$ , where  $\boldsymbol{\tau}_{\eta_\lambda}, \boldsymbol{\tau}_{\eta_u}$  index the prior distributions and may themselves have priors depending on fixed hyperparameters, say,  $\boldsymbol{\alpha}_\eta = (\boldsymbol{\alpha}_{\eta_\lambda}, \boldsymbol{\alpha}_{\eta_u})$ . In practice, we have found it effective to parametrize the distributions  $\lambda_{0p}, f_p$  in a way that permits Gaussian priors. Some example choices for common parametric forms are presented in Appendix B.

The log of the resulting posterior density (up to a constant) is then given by:

$$\begin{aligned} \ell(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\tau}|\mathbf{T}, \boldsymbol{\delta}, \mathbf{Z}) &= \ell(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\eta}, \boldsymbol{\phi}, \mathbf{T}, \boldsymbol{\delta}, \mathbf{Z}) \\ &+ \log \left[ \phi_\lambda^{(\alpha_{\phi_\lambda 1}-1)} (1 - \phi_\lambda)^{(\alpha_{\phi_\lambda 2}-1)} \right] + \log \left[ \phi_u^{(\alpha_{\phi_u 1}-1)} (1 - \phi_u)^{(\alpha_{\phi_u 2}-1)} \right] \end{aligned} \quad (2.15)$$

$$+ \log \pi_\lambda(\boldsymbol{\eta}_\lambda|\boldsymbol{\tau}_{\eta_\lambda}) + \log \pi_u(\boldsymbol{\eta}_u|\boldsymbol{\tau}_{\eta_u}) + \log \pi_\eta(\boldsymbol{\tau}_{\eta_\lambda}, \boldsymbol{\tau}_{\eta_u}|\boldsymbol{\alpha}_\eta) \quad (2.16)$$

where  $\ell(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\eta}, \boldsymbol{\phi}, \mathbf{T}, \boldsymbol{\delta}, \mathbf{Z})$  is analogous to (2.9)–(2.12), with (2.13) and (2.14) substituted for the hazard and frailty curves. The terms in (2.15) correspond to the Beta priors on the weights, and the terms in (2.16) correspond to the priors on the parametric components.

### 2.3.2. Adaptive knot selection

Little attention has been paid in Sections 2.2 and 2.3.1 to the choice of the number of spline knots  $N_\lambda, N_u$  and their positions  $\xi_\lambda, \xi_u$ . Both have a profound effect on the smoothness of the estimated hazard curve or frailty density: including only few widely-spaced knots generally leads to very smooth curves, whereas for multiple knots in close proximity, smoothness has to be enforced by a penalty in order to avoid the risk of overfitting.

In the work of Staudenmayer, Ruppert and Buonaccorsi (2008) and Komárek and Lesaffre (2008), the number of knots is fixed and their positions are distributed evenly over the range of the data. If the underlying curve is smooth, this approach yields excellent results in conjunction with smoothing penalties in the priors of eq. (2.6) and (2.7), provided that the knots and smoothing penalties are well-chosen. However, in the survival analysis setting, such specifications can be more challenging because the hazard function is really only observable through the uncensored event times.

Most forms of penalized smoothing act globally over the range of the data; however, there is often no a priori reason to suspect that all regions of the function(s) being modeled are similarly smooth. One way to relax the imposition of global smoothness in spline smoothing is to permit the data to help determine the “best” choices for both the number and position of the knots. An advantage of the Bayesian approach is that the number and positions of knots may be treated as additional parameters to be estimated. Mallick, Denison and Smith (1999) propose a Bayesian MARS method in which the hazard is a nonparametric function of the covariates, estimated by reversible-jump MCMC (Green, 1995) following Denison, Mallick and Smith (1998). Biller (2000) introduced an approach for automatic knot selection for generalized linear models using natural cubic splines, in which the number of knots and spline weights were chosen by reversible-jump MCMC methods. These procedures make it possible to sample from the posterior of the model set consisting of different numbers and placements of knots, and corresponding different dimensionality of the spline parameters. However, such additional flexibility does not come for free: estimation via MCMC is more complicated in comparison with the fixed knot setting because the total parameter dimension is not fixed a priori, but changes across successive Monte Carlo iterations.

It is possible to extend the spline formulation of Section 2.1 in the manner described above while taking into account the added complexities of the survival model specification and frailty deconvolution. In addition to the steps required to sample from the posterior of the parameters introduced in the preceding sections, the adaptive knot selection procedure consists of three possible moves that affect the hazard and frailty density curves by changing their B-spline basis: the addition of a knot (birth step), the removal of a knot (death step), and the change in position of a knot (move step). We allow the  $\mathbf{N} = (N_\lambda, N_u)$  knots  $\xi = (\xi_\lambda, \xi_u)$  to take positions on a larger, predetermined set of candidate positions, say  $\xi_\lambda^c = \{\xi_{\lambda 1}^c, \dots, \xi_{\lambda M_\lambda}^c\}$ ,  $\xi_u^c = \{\xi_{u 1}^c, \dots, \xi_{u M_u}^c\}$ . These candidate knot specifications constitute a prior on knot placement. For example, the  $M_\lambda$

candidate knots used for the baseline hazard might be selected as quantiles of the observed event times, making data-rich regions more likely to contain knots. During the birth step, an unoccupied candidate knot is added to the knot set; during the death step, an occupied knot is removed; and, during the move step, a knot relocates from its current position to a nearby candidate position.

When selecting the knots adaptively, the smoothness of the curve can be dictated by the number of knots and their positions, with no need for an additional smoothing penalty. In this case, the prior on the number of knots  $\mathbf{N}$  plays a key role in specifying the smoothness. Denison, Mallick and Smith (1998) suggests placing Poisson priors on the number of knots, and simulation experiments conducted by Biller (2000) indicate that this gives good results in the context of nonlinear regression. The Poisson prior is strongly informative, and allows great control over the smoothness of the resulting curve, at the risk of overfitting. In contrast, less informative priors such as a Geometric or Negative Binomial can be used to penalize large numbers of knots and give preference to smoother curves. We consider these and other priors in the supplementary materials.

Let  $\pi(\boldsymbol{\xi}|\mathbf{N})$  and  $\log \pi(\mathbf{N})$  respectively denote the prior distributions selected for knot position and parameter dimension. Then, the joint log-posterior density (up to a constant) is obtained by incorporating this prior information into (2.9)–(2.12); specifically, we have

$$\begin{aligned} \ell(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\xi}, \mathbf{N}|\mathbf{T}, \boldsymbol{\delta}, \mathbf{Z}) \\ = \ell(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}|\mathbf{T}, \boldsymbol{\delta}, \mathbf{Z}, \boldsymbol{\xi}, \mathbf{N}) + \log \pi(\boldsymbol{\xi}|\mathbf{N}) + \log \pi(\mathbf{N}) , \end{aligned} \quad (2.17)$$

where  $\ell(\mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}|\mathbf{T}, \boldsymbol{\delta}, \mathbf{Z}, \boldsymbol{\xi}, \mathbf{N})$  is as in eq. (2.9)–(2.12). A parametric component can be included as well, in the same way as discussed in Section 2.3.1. Further details on the reversible-jump MCMC procedure are provided in Section 3.3.

### 3. Sampling from the posterior distribution

The algorithm used to sample from the joint posterior distribution consists of three types of steps: initialization, parameter updates via Gibbs sampling and Metropolis-Hastings MCMC, and, if desired, adaptive knot selection via reversible-jump MCMC. After initialization, posterior samples of all parameters can be drawn by ordinary and reversible-jump MCMC steps, repeated as long as needed to ensure convergence of the chain and a sufficient number of samples from the posterior. Figure 1 shows the structure of the algorithm. We briefly discuss initialization of the algorithm in Section 3.1 and the Metropolis-Hastings update steps in Section 3.2. The reversible-jump method used for knot selection is discussed in Section 3.3. Expanded versions of these sections are contained in the supplementary materials and in Sharef (2008).

#### 3.1. Initialization

Even though the chain can in principle be initialized at any value, we find that good starting values hasten the convergence of the chain and reduce the risk of

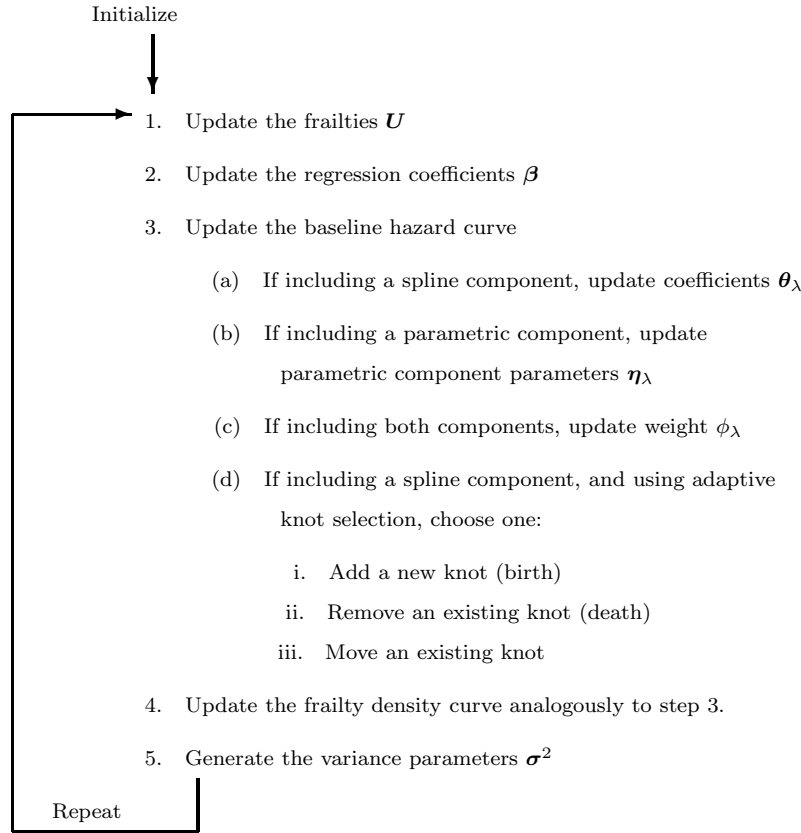


FIG 1. Structure of the sampling algorithm.

numerical problems. Fixing all hyperparameters  $\alpha$  to be 0.01 and all variances  $\sigma$  to be 0.1 results in suitably noninformative priors. We have found that fitting a frequentist proportional hazards model (e.g. via `coxph` in `R`) provides good initial values for both the frailties  $\mathbf{U}$  and regression coefficients  $\beta$ . For the hazard spline, we set the number of knots to  $N_\lambda = \min(\sum_i J_i/4, 35)$  and distribute the knots evenly across the quantiles of the event times.

The frailty density spline is specified on  $N_u = \min(m/4, 35)$  knots distributed evenly across twice the range of the initial values of  $\mathbf{U}$  as determined above, subject to positivity constraints. The spline and parametric coefficients  $\theta_\lambda, \theta_u, \eta_\lambda, \eta_u$  are initialized by maximizing their conditional log-likelihoods given the remaining terms.

### 3.2. Metropolis-Hastings MCMC steps

Once initial values have been obtained, the algorithm proceeds through the MCMC sampling loop of Fig. 1 for a specified number of iterations. Apart from the adaptive knot selection steps, the loop consists of successive Gibbs sampling steps, in which each set of parameters is updated in turn by Metropolis-Hastings. Cluster frailties  $\mathbf{U}$  are updated independently, with candidates generated from gamma distributions centered at their current values. Regression coefficients  $\boldsymbol{\beta}$  and curve coefficients  $\boldsymbol{\theta}_\lambda, \boldsymbol{\eta}_\lambda, \boldsymbol{\eta}_u$  are each updated with normal candidates. The frailty spline parameters  $\boldsymbol{\theta}_u$  are updated pairwise in order to ensure that the frailty density mean remains fixed at 1. That is, when generating a proposal vector, we select two elements, say  $\theta_{ur}$  and  $\theta_{uv}$ , to update; upon generating a new proposal value for  $\theta_{ur}$ , the mean constraint implies the value for  $\theta_{uv}$ . For the weights  $\phi$ , a beta transition kernel is used; lastly, each of the variance parameters in  $\boldsymbol{\sigma}^2$  is sampled from its corresponding conditional posterior Inverse-Gamma distribution. Each of these steps depends on a distinct tuning parameter, set in such a way to maintain an acceptance rate of approximately 25%. Since setting so many tuning parameters by hand is impractical, they can instead be automatically tuned during the burn-in period. Further details on candidate generation and acceptance probabilities may be found in the supplementary materials and also in Sharef (2008).

### 3.3. Reversible-jump MCMC for adaptive knot selection

Throughout the steps discussed in Section 3.2, the number of knots in the model, hence dimension of the spline parameters  $\boldsymbol{\theta}_\lambda, \boldsymbol{\theta}_u$ , has remained fixed. In order to enable adaptive knot selection, we not only allow knots to move, but also permit changes in dimension, such as adding a knot (birth step) or deleting a knot (death step). The basic procedure is identical for both the hazard function and frailty density; hence, we discuss the procedure below in general terms only, omitting subscripts that identify the parameters as referring to either curve.

As indicated above, adaptive knot selection requires three types of steps: the “move” step, in which the position of a single knot is changed to some new point between its neighbor knots; the “birth” step, in which a new knot is added after a randomly chosen knot and the dimension of the spline parameter  $\boldsymbol{\theta}$  increases; and, the “death” step, in which a randomly chosen knot is removed and the dimension of the parameter decreases. The move step requires no dimension change and a standard Metropolis-Hastings update is sufficient. However, the death and birth steps involve a change in model dimension; hence, we use reversible-jump MCMC subject to a “dimension-matching” constraint (Green, 1995). Typically, transitions between a model indexed by a parameter set  $\boldsymbol{\theta}$  of dimension  $k$  and a candidate model indexed by parameters  $\tilde{\boldsymbol{\theta}}$  of dimension  $\tilde{k}$  are accomplished by generating  $m$  uniform random numbers  $\mathbf{u}$  and computing the candidate by a deterministic function  $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{u})$ . For the reverse move, one generates  $\tilde{m}$  random numbers  $\tilde{\mathbf{u}}$  and computes the candidate as  $\boldsymbol{\theta} = \boldsymbol{\theta}(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{u}})$ . To

ensure reversibility, the mapping between  $(\boldsymbol{\theta}, \mathbf{u})$  and  $(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{u}})$  must be bijective, and in particular, the dimension-matching constraint  $m + k = \tilde{m} + \tilde{k}$  must hold.

Following Denison, Mallick and Smith (1998), at each iteration we choose randomly whether to execute a birth, death, or move step. Given  $N = n$ , the probability of taking birth, death or move step is respectively

$$b_n = c \min \left\{ 1, \frac{\pi_N(n+1)}{\pi_N(n)} \right\}, \quad d_n = c \min \left\{ 1, \frac{\pi_N(n-1)}{\pi_N(n)} \right\}, \quad m_n = 1 - b_n - d_n,$$

where the constant  $c$  controls the rate of dimension-changing steps and  $b_n \pi_N(n) = d_{n+1} \pi_N(n+1)$ . As in Denison, Mallick and Smith (1998), we take  $c = 0.4$ . We provide further details on the move step in Section 3.3.1, the birth step in Section 3.3.2, and the death step in Section 3.3.3.

### 3.3.1. Knot position change (move step)

In the move step, a single knot position  $\xi_k$  to be moved is chosen uniformly from the set of interior knots, and changed to a random new candidate position located between its neighboring knots. That is, the candidate knot position  $\tilde{\xi}_k$  is selected uniformly from the set of candidate locations  $\xi^c \in \boldsymbol{\xi}^c$  such that  $\xi_{k-1} < \xi^c < \xi_{k+1}$ . Since the prior on the knot positions is discrete uniform over the set of candidate knots, the prior probabilities for knots  $\boldsymbol{\xi}$  and the candidate  $\tilde{\boldsymbol{\xi}}$  are identical. Since no dimension change is required, the new knot positions are accepted with a Metropolis-Hastings probability. The spline parameters  $\boldsymbol{\theta}$  remain unchanged.

### 3.3.2. Knot addition (birth step)

In the birth move, a random unoccupied candidate knot  $\xi^c \in \boldsymbol{\xi}^c$  is chosen to be added to the current set of knots  $\boldsymbol{\xi}$ , of length  $N$ . Denote by  $k$  the interval of the current knot set containing  $\xi^c$ , so that  $\xi_k < \xi^c < \xi_{k+1}$ . The new candidate knot set is then given by  $\tilde{\boldsymbol{\xi}} = \{\xi_1, \dots, \xi_k, \xi^c, \xi_{k+1}, \dots, \xi_N\}$ , of length  $\tilde{N} = N + 1$ .

The set of spline coefficients  $\boldsymbol{\theta}$  of length  $K$  must be updated to a candidate set  $\tilde{\boldsymbol{\theta}}$  of length  $\tilde{K} = K + 1$ . There are simple rules for non-destructively inserting a new knot into a B-spline function (de Boor, 2001), but using these directly would violate the reversibility and dimension-matching constraint between the birth and death moves mentioned earlier. Since the birth move begins in a model of dimension  $K$  and its reverse begins at dimension  $K + 1$ , we need to generate an additional random number for the birth move. Intuitively, since removing a knot is a destructive procedure and may cause the shape of the curve to change, we must during the birth move be able to generate the set of curves that would reduce to the original curve upon removal of the new knot.

To do this, we compute the candidate spline parameters  $\tilde{\boldsymbol{\theta}}$  for inserting a knot  $\xi^c \in (\xi_k, \xi_{k+1})$  as follows:

$$\tilde{\theta}_j = \begin{cases} \theta_j & \text{if } j \leq k+1 \\ \theta_{j-1} & \text{if } j > k+Q \\ \log(r_j e^{\theta_j} + (1-r_j)e^{\theta_{j-1}}) & \text{if } k+1 < j < k+Q \\ \log(ue^{\theta_j} + (1-u)e^{\theta_{j-1}}) & \text{if } j = k+Q \end{cases} \quad (3.1)$$

where  $r_j = (\xi^c - \xi_{j-Q})/(\xi_{j-1} - \xi_{j-Q})$ , and  $u \sim U(0, 1)$ . These rules correspond to the deterministic rules in de Boor (2001), except that the parameter  $\tilde{\theta}_{k+Q}$  is perturbed by a random amount, rather than by the knot ratio  $r_{k+Q}$ . The candidate number of knots  $\tilde{N}$  and spline parameters  $\tilde{\theta}$  are then accepted with probability

$$\rho = \min\{1, R_L \cdot R_P \cdot R_T \cdot |J|\}, \quad (3.2)$$

where  $R_L$  is the relevant likelihood ratio (see (2.17)),  $R_P$  is the prior ratio encompassing the priors on the number of knots and knot positions,  $R_T$  is the transition ratio and  $J$  is the Jacobian of the transformation in (3.1). Expressions for these terms may be found in the supplementary materials.

### 3.3.3. Knot deletion (death step)

In the death step, a single knot  $\xi_k$  is chosen uniformly from the set of knots  $\xi$  to be removed. The candidate knot set for the next iteration is then  $\tilde{\xi} = \{\xi_1, \dots, \xi_{k-1}, \xi_{k+1}, \dots, \xi_N\}$ . The spline parameters are correspondingly adjusted by the inverse of the transformation in (3.1), that is, by deleting  $\theta_{k+Q-1}$  and adjusting the remaining parameters as

$$\tilde{\theta}_j = \begin{cases} \theta_j & \text{if } j < k+1 \\ \theta_{j+1} & \text{if } j \geq k+Q-1 \\ \log\left(\frac{1}{r_j}e^{\theta_j} - \frac{1-r_j}{r_j}e^{\theta_{j-1}}\right) & \text{if } k+1 \leq j < k+Q-1 \end{cases}$$

Because the birth and death moves are symmetrically defined, the likelihood ratio, prior ratio, transition ratio and Jacobian determinant are the inverses of those in eq. (3.2).

## 4. Simulation studies

We implemented the methodology described in Section 3 in the R package **splinesurv**. In order to establish the performance and flexibility of the method, we conducted simulation studies under a variety of settings that intend to investigate the capacity of the method to correctly identify the form of the underlying baseline hazard and frailty density, for different numbers of clusters and cluster sizes. Furthermore, we desired to show that the estimated regression coefficients  $\beta$  and frailty variance parameter  $\nu^2$  can be accurately estimated. We remark here that  $\nu^2$  is not an explicit model parameter but rather a functional of the frailty density; for example, under (2.4),  $\nu^2 = \nu^2(\theta_u)$ , where

$$\nu^2(\theta_u) = \int_{x \geq 0} (x-1)^2 f(x|\theta_u) dx; \quad (4.1)$$



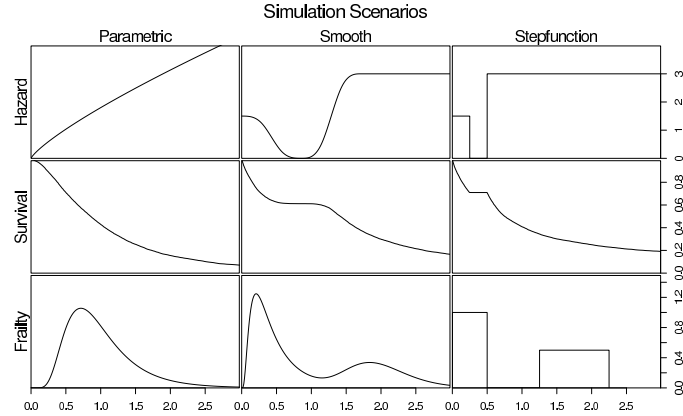


FIG 2. “True” baseline hazard curves, survival functions and frailty densities used for generating simulated data in each of the three simulation scenarios.

alternatively, one may replace (2.4) with (2.14). We focus on this particular summary since it is frequently of interest in analyses utilizing a parametric frailty model.

We consider three scenarios within which to test the method, differing in the form of the “true” baseline hazard and frailty density used to generate simulated data. The first, referred to as the “Parametric” scenario, is characterized by a Weibull hazard and a lognormal frailty density. In the “Smooth” scenario, the baseline hazard is a smooth curve not easily described by typical parametric forms, and the frailty density is a mixture of two lognormal distributions. In the “Stepfunction” scenario, the baseline hazard is a discontinuous step function, and the frailty distribution is a mixture of uniform densities. Figure 2 contains plots of the hazard, survival and frailty in each of the three scenarios.

For purposes of the simulation, a replication consists of first generating frailties  $U_i$ ,  $i = 1 \dots m$  from the scenario’s frailty density. A single covariate is generated for each subject as  $Z \sim N(0, 1)$ . The single regression coefficient is fixed at  $\beta = 1$ . Given the frailty and covariate, event times can then be generated using the baseline hazard for the scenario. Censoring times are independently generated from a Weibull hazard with parameters chosen for each scenario to yield approximately a 20% censoring rate. The simulated sample generated in this way can then be fit using the `splinesurv` package.

#### 4.1. Curve fitting performance

To explore the effects of sample size on the quality of the curve fits, we first simulate a single dataset for various sample sizes, under each scenario, and explore the effect of different model specifications. We limit ourselves to four sample sizes for each scenario, setting the number of clusters to either  $m = 10$  or  $m = 500$ , and the cluster size to either  $J_i = 10$  or  $J_i = 500$ ,  $i = 1 \dots m$ .

We also include a sample size and censoring profile matching the data to be analyzed in Sec. 5

The methodology is very flexible, offering a range of choices of penalty functions, parametric distributions, prior parameters, and the option of adaptive knot selection. For brevity, we only select one model specification for purposes of demonstrating curve-fitting here. For both the hazard and frailty, we include a spline component only, using a simple Gaussian prior on the spline parameters (corresponding to the penalty function in Section A.1). We allow for adaptive knot selection with a truncated Poisson prior on the number of knots, with means  $\mu_\lambda = \mu_u = 10$  and a maximum of 35 knots, and 100 candidate knots distributed uniformly over the range of the data. Each fit was run for a 2000-iteration burn-in, during which tuning parameters were chosen adaptively to ensure approximately a 25% parameter acceptance rate, followed by 3000 iterations used for constructing posterior estimates.

The results are shown in Figure 3, where we report posterior means and 95% pointwise credible intervals, and suggest that the methodology functions as intended: in each scenario, the fitted models capture the features of the underlying hazard and frailty curves with sufficiently large samples. The number of clusters appears to have a more immediate effect on the quality of the fit than does the cluster size, especially for the frailty density. In order to obtain an accurate estimate of the frailty density, a large number of clusters is required, but these clusters need not be large. With few, large clusters, the form of the hazard can be identified, while the frailty distribution cannot. Hazard estimates in the Stepfunction scenario display sharp spikes at the points of discontinuity—this is an artifact caused by the use of cubic splines in a scenario where linear splines would have been better able to capture the discontinuity.

Following these single dataset analyses, we repeated this process 1000 times. Averaging the posterior mean estimates over multiple replications shows that the baseline hazard curve is estimated with great accuracy and only minimal smoothing bias (see Fig. 4). If the number of clusters is large, the frailty density can be well-estimated also. Further such simulation results (not shown here) indicate that fixed-knot penalized splines perform well in the Parametric and Smooth scenarios, but do quite poorly in the Stepfunction scenario, as the sharp trough cannot be captured without significant smoothing error. Further, the inclusion of a correctly specified parametric component improves curve-fitting performance in the Parametric scenario, and does not significantly affect the other scenarios.

#### 4.2. Parameter estimation performance

In order to establish the ability of the procedure to estimate the regression parameter and the frailty variance, we conduct a simulation study at smaller sample sizes. Sample sizes under all three scenarios are limited to 10, 50 and 500 clusters of size 10 or 50 or 500, excluding the largest combination. We further introduce a sample with the same cluster number and sizes, and censoring

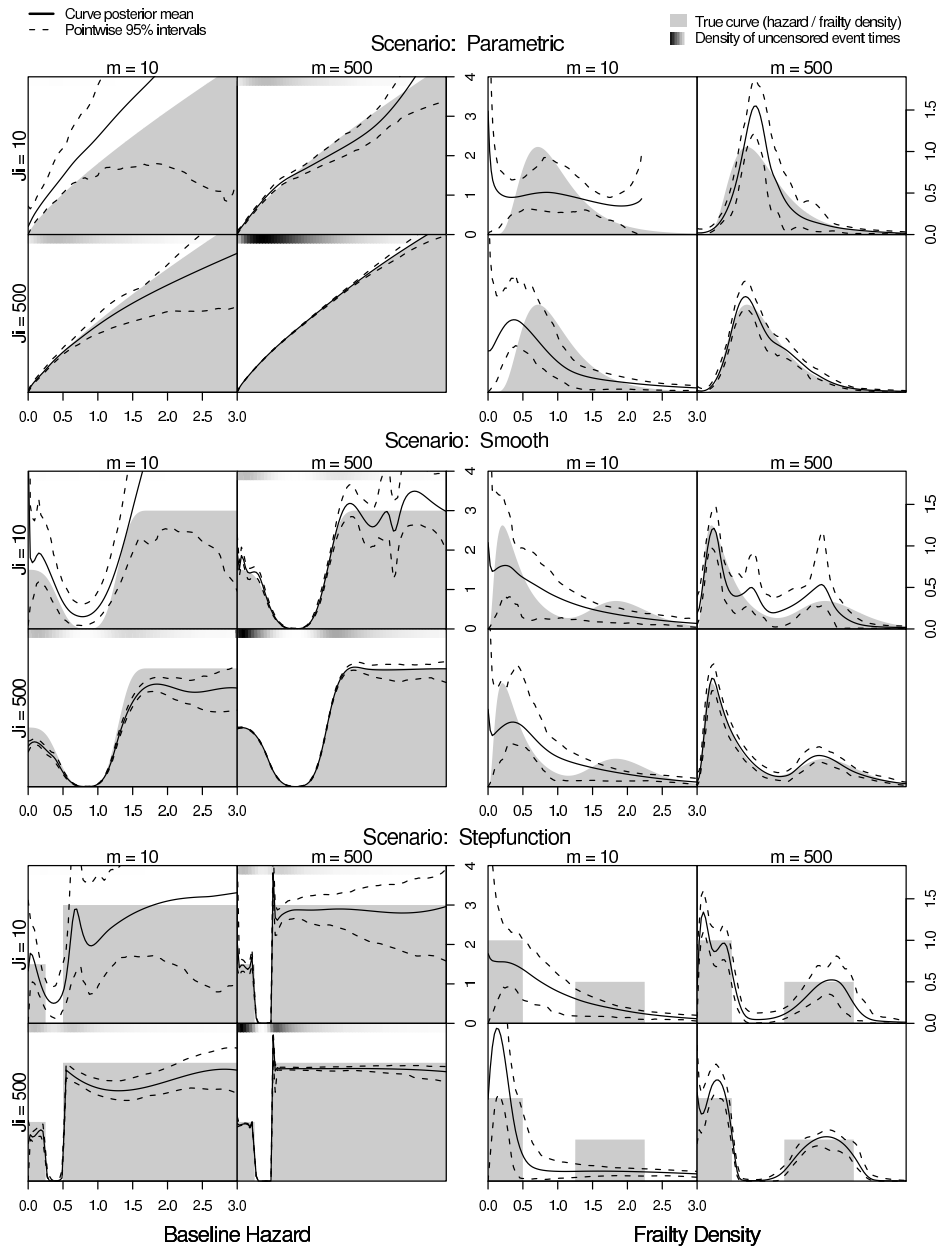


FIG 3. Baseline hazard and frailty density curve fitting results of a single replication conducted under each of the three scenarios, for different sample sizes.

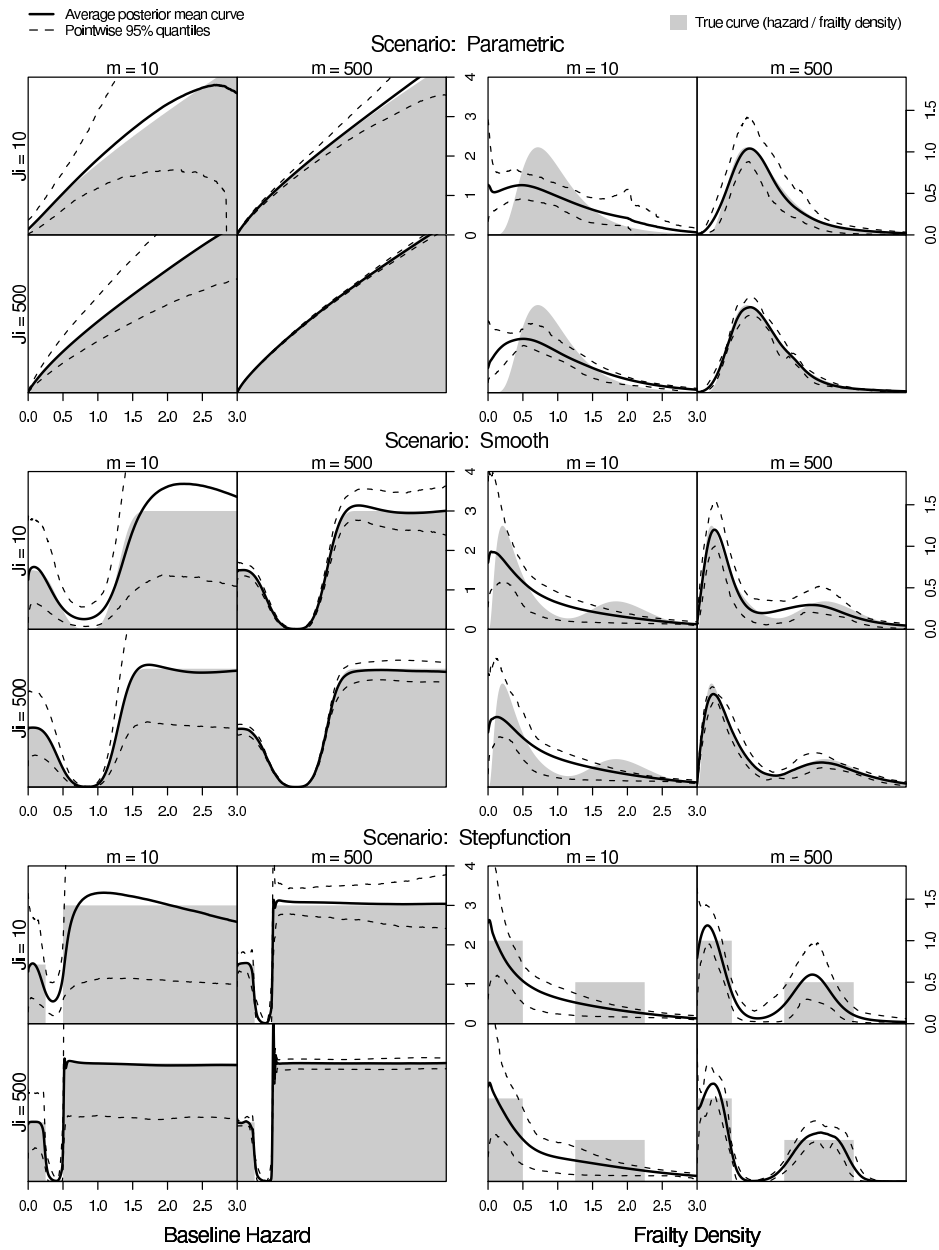


FIG 4. Baseline hazard and frailty density estimates, aggregated over 1000 replications (10 for the largest sample) for each of the three scenarios, for different sample sizes.

characteristics similar to the data set to be analyzed in sec. 5. The scenarios are specified as before.

We consider six spline model specifications: The first two are fourth order spline-only models with adaptively chosen knots and Poisson(10) or Poisson(20) priors respectively on the number of knots, as described for Figure 3. The third and fourth add parametric components, consisting of a Weibull baseline and a lognormal frailty curve, with a Beta(1, 2) prior on the weight, thus giving slight preference to the parametric component. We note here that these parametric components are correctly specified in the Parametric scenario but not otherwise. The fifth is a fixed-knot penalized spline model specified according to Section 2.1, with equally spaced knots and a penalty on the integrated squared second derivative, following Section A.3. The sixth is similar, but penalizes the squared second differences between the parameters, as per Section A.2. Penalized spline fits can be sensitive to the choice of hyperparameters; these were chosen here so as to give reasonably smooth curves in several test scenarios; we intentionally did not choose the “best” settings, but instead tried to select parameters as one might do if the underlying curve were unknown.

In Bayesian estimation by MCMC, the collection of posterior samples contains a great deal of useful information. However, for the purposes of this simulation study, and to enable comparison with frequentist methodology, we construct point estimators from these posterior samples and summarize their frequentist performance. Natural estimators for the regression coefficient  $\beta$  include the mean and median of its marginal posterior distribution, estimated for each simulated dataset from the corresponding MCMC samples from the joint posterior distribution. Similarly, for the frailty variance  $\nu^2$ , one may utilize the posterior mean or median of (4.1), obtained for each simulated dataset using the corresponding MCMC samples from the posterior distribution of the parameters defining either (2.4) or (2.14).

Tables 1 and 2 respectively contain the simulated biases of the posterior mean and median point estimators based on 1000 simulations for each of the spline model specifications. For comparison, we also fit a (misspecified) gamma frailty model as described in Therneau and Grambsch (2000) using the R routine `coxph`. In general, the results show that the proposed Bayesian modeling scheme yields good estimates of the regression coefficient, with the similarity between means and medians reflecting symmetry in the posterior. With few exceptions, the biases are acceptably small and similar to the regression coefficients estimated under the gamma frailty model.

When additionally accounting for estimation of the frailty variance, we see that the adaptive methods perform very well in the Parametric setting; here, increasing the prior mean number of knots also has no clear effect, as evidenced by the comparable performance of the Adapt(10) and Adapt(20) methods. In fact, most of the methods perform quite well in the Parametric setting across the range of sample sizes considered, a notable exception being the Penalized (2<sup>nd</sup> diff.) prior when estimating  $\nu^2$ . The story is somewhat less consistent outside of the Parametric setting, where the Penalized (2<sup>nd</sup> deriv.) prior arguably exhibits some robustness and turns in the best overall performance. The Penal-

TABLE 1  
 Percentage bias of posterior means of the regression coefficient and frailty variance in six model specifications, and a gamma frailty model. Results are based on 1000 replications

Method	$m$	$J_i$	Censor	Parametric		Smooth		Stepfunction			
				Bias %	$\nu^2$	Bias %	$\nu^2$	Bias %	$\nu^2$		
Adapt(10)	10	10	0.20	-1.51	5.82	3.32	41.87	1.25	63.08		
		50	0.20	-0.78	3.28	0.32	31.44	0.57	59.66		
		500	0.20	0.22	-0.61	0.17	27.26	0.15	60.02		
	50	10	0.20	0.79	-0.31	1.60	20.95	1.03	39.49		
		50	0.20	0.13	1.95	0.45	10.33	0.03	29.60		
		500	10	0.20	0.28	-14.05	0.75	12.22	0.59	17.13	
	31	6	0.40	-0.11	-4.22	2.39	37.26	0.82	65.44		
		Adapt(20)	10	10	0.20	-2.17	2.16	3.02	55.01	-2.46	75.60
				50	0.20	-0.35	6.36	0.90	40.15	0.61	65.23
500	0.20			0.00	8.58	0.17	37.28	0.13	67.22		
50	10		0.20	1.19	-1.34	0.39	24.78	1.50	53.26		
	50		0.20	0.44	2.09	0.32	14.96	0.06	34.16		
	500		10	0.20	-0.09	-12.40	0.38	11.41	0.69	16.54	
31	6		0.40	-1.27	-0.27	3.35	50.79	0.60	72.00		
	Adapt(10)+Par		10	10	0.20	0.87	-10.30	-1.02	36.39	-2.84	82.45
				50	0.20	-0.16	-5.00	-0.21	35.46	-0.55	77.80
500		0.20		0.05	-6.76	-0.06	34.89	-0.11	83.79		
50		10	0.20	-0.59	-10.18	0.21	33.41	0.00	77.88		
		50	0.20	-0.27	-10.95	-0.01	26.05	-0.00	60.69		
		500	10	0.20	-0.25	-21.13	0.63	19.73	0.72	28.82	
31		6	0.40	-2.57	-6.95	-0.16	43.41	-1.92	91.32		
		Adapt(20)+Par	10	10	0.20	-1.58	-2.67	-0.40	38.59	-5.66	81.94
				50	0.20	-0.79	-1.11	0.30	38.22	-0.47	91.63
500	0.20			-0.14	-5.70	0.06	38.83	0.00	89.04		
50	10		0.20	0.05	-10.69	0.63	33.63	0.61	80.34		
	50		0.20	0.02	-11.38	0.36	26.55	0.22	59.99		
	500		10	0.20	0.07	-20.58	0.48	19.61	0.58	27.57	
31	6		0.40	-2.53	-4.56	0.85	41.90	-1.42	89.30		
	Pen. (2 <sup>nd</sup> deriv.)		10	10	0.20	4.44	-14.32	-16.00	-15.12	-11.25	3.03
				50	0.20	-0.98	-12.98	-6.08	-0.86	-6.01	18.93
500		0.20		1.23	-4.99	-0.41	12.26	-1.36	20.39		
50		10	0.20	-0.20	-10.05	-6.07	-1.17	-5.93	3.66		
		50	0.20	0.60	1.86	-1.16	3.46	-1.74	8.16		
		500	10	0.20	1.09	-4.61	0.04	2.54	-0.98	9.11	
31		6	0.40	6.90	-12.76	-11.22	-6.31	-6.70	10.45		
		Pen. (2 <sup>nd</sup> diff.)	10	10	0.20	2.05	-39.31	0.40	-15.29	2.68	8.51
				50	0.20	0.86	-30.45	0.19	-8.08	-1.30	4.51
500	0.20			-0.19	-28.79	-0.17	-6.22	-0.63	6.07		
50	10		0.20	-1.38	-44.00	-0.35	-13.12	-1.24	2.82		
	50		0.20	-0.74	-31.34	-0.15	-6.85	-0.08	2.41		
	500		10	0.20	-0.44	-15.50	0.02	0.55	-0.43	4.02	
31	6		0.40	-3.91	-59.44	-3.43	-23.60	-1.72	1.31		
	Cox: gamma		10	10	0.20	-0.89	-44.26	0.02	7.73	-2.46	52.53
				50	0.20	-1.09	-12.91	-0.23	29.70	-0.64	74.10
500		0.20		0.29	33.98	-0.02	47.93	-0.28	93.92		
50		10	0.20	0.52	-35.16	0.02	9.70	0.29	75.17		
		50	0.20	1.03	-3.51	0.51	34.70	0.25	69.11		
		500	10	0.20	-0.89	-37.63	0.08	14.27	0.06	70.07	
31		6	0.40	2.27	-45.01	0.78	9.50	-1.86	48.16		

ized (2<sup>nd</sup> diff.) prior is observed to perform especially well in the Stepfunction (i.e., nonsmooth) setting and acceptably well in the Smooth setting; the adaptive methods also perform well across all settings with larger total sample sizes. The adaptive methods that include a parametric component perform admirably in estimating the regression coefficients in all cases. These also do very well in estimating the frailty variance in the Parametric setting (i.e., where both para-

TABLE 2  
 Percentage bias of posterior medians of the regression coefficient and frailty variance in six model specifications. Results are based on 1000 replications

Method	$m$	$J_i$	Censor	Parametric Bias %		Smooth Bias %		Stepfunction Bias %			
				$\beta$	$\nu^2$	$\beta$	$\nu^2$	$\beta$	$\nu^2$		
Adapt(10)	10	10	0.20	-1.76	0.60	3.16	36.80	1.07	57.97		
		50	0.20	-0.88	-2.96	0.25	24.36	0.48	55.73		
		500	0.20	0.19	-6.14	0.13	20.92	0.09	57.08		
	50	10	0.20	0.70	-3.61	1.54	16.61	0.99	34.31		
		50	0.20	0.10	-0.30	0.39	6.53	-0.02	26.06		
		500	10	0.20	0.25	-14.75	0.69	11.73	0.55	16.51	
	31	6	0.40	-0.34	-8.98	2.33	31.46	0.70	59.28		
		Adapt(20)	10	10	0.20	-2.46	-0.21	2.87	53.03	-2.58	73.04
				50	0.20	-0.42	1.09	0.86	35.43	0.57	62.17
500	0.20			-0.04	3.66	0.11	32.49	0.08	65.74		
50	10		0.20	1.11	-4.08	0.31	19.85	1.45	49.84		
	50		0.20	0.40	0.55	0.26	11.11	0.01	31.31		
	500		10	0.20	-0.11	-12.87	0.34	10.83	0.65	15.82	
31	6		0.40	-1.45	-2.91	3.27	48.67	0.48	68.14		
	Adapt(10)+Par		10	10	0.20	0.74	-13.33	-1.15	24.55	-2.90	47.54
				50	0.20	-0.18	-4.71	-0.25	21.59	-0.58	45.77
500		0.20		0.04	-3.81	-0.11	22.94	-0.16	52.66		
50		10	0.20	-0.63	-6.27	0.14	17.27	-0.04	37.08		
		50	0.20	-0.29	0.72	-0.06	6.42	-0.06	30.46		
		500	10	0.20	-0.26	-16.74	0.58	15.45	0.69	21.96	
31		6	0.40	-2.74	-16.37	-0.29	22.60	-2.10	52.45		
		Adapt(20)+Par	10	10	0.20	-1.66	-2.23	-0.54	33.79	-5.79	59.84
				50	0.20	-0.83	2.96	0.27	33.49	-0.52	70.30
500	0.20			-0.15	-0.84	0.01	34.28	-0.04	64.30		
50	10		0.20	0.02	-7.24	0.56	17.28	0.52	44.84		
	50		0.20	0.01	-1.76	0.30	9.59	0.17	35.44		
	500		10	0.20	0.06	-16.16	0.43	15.69	0.54	20.66	
31	6		0.40	-2.57	-13.05	0.72	28.71	-1.52	56.56		
	Pen. (2 <sup>nd</sup> deriv.)		10	10	0.20	4.39	-15.37	-15.99	-18.84	-11.11	1.50
				50	0.20	-0.93	-14.99	-6.04	-1.93	-5.95	19.81
500		0.20		1.23	-6.32	-0.42	12.32	-1.36	20.28		
50		10	0.20	-0.17	-12.12	-6.07	-3.85	-5.91	0.21		
		50	0.20	0.60	0.30	-1.17	1.18	-1.72	5.20		
		500	10	0.20	1.09	-4.89	0.05	2.09	-0.97	8.48	
31		6	0.40	6.81	-15.03	-11.16	-9.08	-6.65	7.24		
		Pen. (2 <sup>nd</sup> diff.)	10	10	0.20	1.98	-43.40	0.30	-18.87	2.55	5.41
				50	0.20	0.82	-34.79	0.18	-10.32	-1.30	3.00
500	0.20			-0.19	-32.63	-0.16	-8.32	-0.63	4.20		
50	10		0.20	-1.43	-45.21	-0.38	-14.17	-1.27	2.31		
	50		0.20	-0.76	-32.72	-0.15	-7.36	-0.08	1.99		
	500		10	0.20	-0.44	-15.87	0.02	0.12	-0.43	3.55	
31	6		0.40	-4.04	-62.70	-3.56	-26.44	-1.87	-1.70		

metric components are correctly specified). Outside of this setting, the degree of smoothness and sample / cluster size have a clear impact on performance.

In general, the results suggest that the Penalized (2<sup>nd</sup> deriv.) prior is the best overall choice unless one has strong prior information to suggest that the baseline hazard and frailty distributions are particularly smooth (in which case adaptive methods with or without a parametric start perform commendably well) or particularly nonsmooth (in which case Penalized (2<sup>nd</sup> diff.) is perhaps the best choice). The frequentist performance of the posterior mean and median tend

to be comparable for estimating regression coefficients; the posterior median exhibits somewhat less bias in comparison with the mean for estimating the frailty variance, with the greatest benefits being observed in the Smooth and Stepfunction settings when used in combination with adaptive knot selection.

As expected, the results for the gamma frailty model in Table 1 exhibit good performance for estimating the regression coefficient in all settings, despite a misspecified frailty distribution; however, relatively poor performance is observed for estimating the frailty variance. In our opinion, the results suggest that the use of the Penalized (2<sup>nd</sup> deriv.) prior creates an acceptable and generally robust tradeoff, with the potential for mildly increased bias in the regression coefficients being offset by substantial reductions in the bias of the estimated frailty variance.

## 5. Example

### 5.1. Congestive heart failure data description

We consider data from a study conducted in a 487-bed, not-for-profit community hospital located in southeast Michigan. The study population consisted of patients with either systolic or diastolic heart failure, and was originally identified for the purposes of a randomized, controlled clinical trial designed to compare a pro-active case management strategy to standard care on all-cause re-hospitalizations. A planned secondary analysis was to determine prognostic factors for readmission or mortality.

Patients were eligible for the study if they were hospitalized on an internal or family medicine service between October 29, 2002 and September 20, 2003 and received intravenous diuretics to treat possible heart failure. Intervention patients were assessed by a cardiology nurse practitioner who developed a protocol-driven discharge plan that could include telemanagement, an outpatient nurse-run heart failure clinic, or usual care. All control patients were managed by the usual discharge planning activities of hospital staff. Of the 440 patients enrolled in the study, 17 died during the index hospitalization and were removed from the sample, resulting in a cohort of 423 patients. Unfortunately, approximately one-half of the patients assigned to the intervention arm were discharged prior to receiving the complete intervention. Using an intent-to-treat analysis, (e.g. Fisher *et al.*, 1990), no difference between the intervention or control groups was found for the outcome of all-cause subsequent hospitalizations or emergency department encounters.

Below, we proceed to re-analyze the data, defining the event of interest as a patient's first rehospitalization or death during the 180 day period following the index hospitalization. Of the 423 patients, 257 such events were observed, of which 233 are rehospitalizations and the remainder are deaths. The remaining 39.2% of patients are considered to be censored as of the end of followup. Patients are clustered into 31 groups by their attending physician, ranging in size between 1 and 80 patients, with mean and median cluster sizes of 14 and



TABLE 3  
Covariates and basic descriptive statistics for the congestive heart failure data

Name	Description	Mean	Median	SD
hxsumINPTorER	Prior hospitalizations and ED visits (count)	0.88	0.00	1.31
minHb	Minimum hemoglobin	10.79	10.80	2.07
LN_lastCREAT	Last creatinine log	0.26	0.22	0.45
LN_maxGLU	Maximum glucose log	5.12	5.04	0.41
minPLTSl50k	Indicator: minimum platelet count $\geq 10^5$	0.02	0.00	0.13
lastPOTASgt5	Indicator: last potassium $>5$	0.04	0.00	0.19
itoECF	Indicator: discharged to nursing home	0.19	0.00	0.39
ejectionpctcon	Cardiac ejection fraction	43.66	45.00	16.47
dcbeta	Indicator: beta-blockers	0.54	1.00	0.50
dcaceiorarb	Indicator: ACE Inhibitors	0.61	1.00	0.49

5.5 respectively. A wide range of explanatory variable data are available for each patient. Since treatment was not administered to half of the intervention group patients, treatment group membership is excluded from the set of covariates. Covariates with more than 5% missing values were not considered for the purposes of this analysis; for all remaining covariates, we imputed any missing covariate values by their respective median values. A subset of these remaining covariates has been selected using a combination of stepwise automated procedures and consultations with study clinicians. These covariates, along with basic descriptive statistics, are summarized in Table 3.

## 5.2. Model selection and fitting

In the absence of external information on which to base a model choice, we fit all six model specifications presented in Sec. 4.2. Prior to analysis, all covariates were centered and standardized; we found empirically that doing so improved the mixing properties of the MCMC procedure. The sampling chain is run for 50,000 iterations, discarding the first 20,000 iterates as burn-in and thinning the chain to every 10th sample. In order to provide a measure of model discrimination and means for selection, we used a modified version of the original Deviance Information Criterion (DIC; Spiegelhalter *et al.*, 2002). In particular, we computed

$$DIC = \bar{D} + \frac{V}{2}, \quad (5.1)$$

where  $\bar{D}$  and  $V$  are respectively the sample mean and variance of deviance calculated at each MCMC step. Here,  $V/2$  is used as an estimator of the effective number of parameters  $p_D$  because, unlike the usual estimator, it is guaranteed to be non-negative; see (Gelman *et al.*, 2004, p. 182). Estimated DIC values are shown Table 4. Consistent with simulation results, these results indicate that a model fitting penalized splines to the hazard and frailty density using a prior that reflects a second-derivative-type penalty is a good choice.

We focus our attention below on the indicated model; results pertinent to this model fit are summarized in Table 5 and Figure 5. Recall the hazard and frailty

TABLE 4  
Deviance Information Criterion estimates for six model specifications, fitted to the congestive heart failure data

Model	DIC
Adapt(10)	3487.1
Adapt(20)	3869.5
Adapt(10)+Par	3424.7
Adapt(20)+Par	3604.59
Pen. (2 <sup>nd</sup> deriv.)	2807.7
Pen. (2 <sup>nd</sup> diff.)	3033.5

TABLE 5  
Posterior means (PM) and 95% credible intervals of regression coefficients and frailty variance, for models using penalized splines and adaptive splines, fitted to congestive heart failure data, compared to gamma and lognormal frailty model estimates from coxph

Covariate	Pen. (2 <sup>nd</sup> deriv.)			Adapt(10)		
	PM	2.5%	97.5%	PM	2.5%	97.5%
hxsumINPTorER	0.324	0.197	0.444	0.245	0.127	0.354
minHb	-0.256	-0.409	-0.107	-0.199	-0.335	-0.067
LN_lastCREAT	0.196	0.036	0.357	0.157	0.017	0.303
LN_maxGLU	0.120	-0.031	0.264	0.113	-0.023	0.244
minPLTSlt50k	0.144	0.025	0.252	0.120	0.008	0.224
lastPOTASgt5	0.138	0.010	0.255	0.079	-0.041	0.191
itoECF	0.112	-0.017	0.240	0.104	-0.013	0.219
ejectionpctcon	-0.129	-0.289	0.024	-0.104	-0.238	0.027
dcbeta	-0.057	-0.200	0.085	-0.039	-0.167	0.090
dcaceiorarb	-0.052	-0.197	0.092	-0.034	-0.156	0.093
minHb:LN_lastCREAT	0.100	-0.044	0.259	0.167	0.032	0.308
Frailty Variance	0.406	0.216	0.692	0.716	0.176	1.533

Covariate	Cox: gamma			Cox: lognormal		
	Est	SD	Pval	Est	SD	Pval
hxsumINPTorER	0.248	0.056	0.000	0.249	0.056	0.000
minHb	-0.193	0.069	0.005	-0.195	0.070	0.005
LN_lastCREAT	0.236	0.074	0.001	0.235	0.074	0.002
LN_maxGLU	0.122	0.067	0.068	0.123	0.067	0.066
minPLTSlt50k	0.138	0.051	0.007	0.138	0.051	0.007
lastPOTASgt5	0.122	0.055	0.026	0.120	0.055	0.028
itoECF	0.129	0.060	0.033	0.129	0.060	0.033
ejectionpctcon	-0.117	0.071	0.098	-0.119	0.071	0.094
dcbeta	-0.055	0.065	0.398	-0.055	0.065	0.400
dcaceiorarb	-0.045	0.066	0.502	-0.043	0.066	0.517
minHb:LN_lastCREAT	0.189	0.073	0.010	0.190	0.073	0.009
Frailty Variance	0.000			0.001		

are specified following Sections 2.1 using fourth-order (cubic) splines. Table 5 shows the posterior mean and 95% posterior intervals for the covariate effects and the variance of the random effects in the first three columns. We observe substantial agreement in the signs and magnitudes of the regression coefficients in comparison with both the gamma and lognormal frailty models fit using coxph and summarized in the lower half of Table 5. For comparison, we also

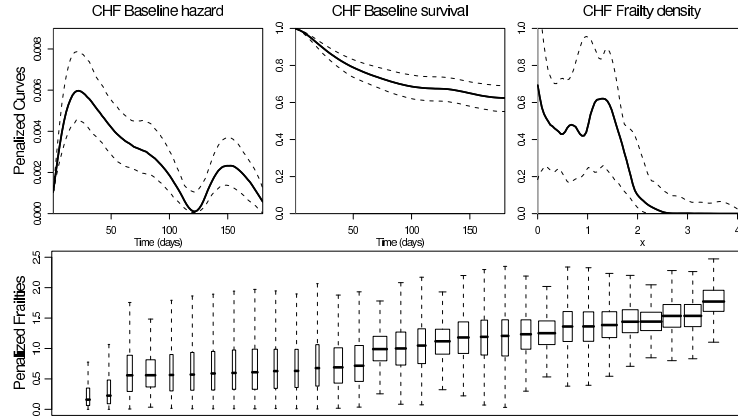


FIG 5. Upper panel: Hazard, survival and frailty density estimates, and 95% pointwise posterior intervals for the congestive heart failure data. The baseline represents a patient with average covariates. Lower panel: Boxplots of the marginal posterior distributions for each of 31 cluster-level frailties, sorted in order of ascending posterior mean. Box width indicates the cluster size.

provide the regression results and frailty variance obtained under the Adapt(10) spline model specification. Comparable results are again observed for the regression parameters, with the frailty variance estimate having a substantially larger posterior mean and variation.

Both of the spline-based models suggest the existence of a frailty effect, whereas the frailty variance is effectively estimated to be zero in both the gamma and lognormal frailty models. Based on the results summarized in Section 4, care is clearly required when interpreting these point estimates; all variance estimators exhibit some bias. For example, with the indicated sample size and under the Smooth setting (e.g., Table 1), we see a positive bias for the Adapt(10) setting and negative bias for the Pen. (2<sup>nd</sup> deriv.) setting, suggesting that the posterior mean of the frailty variance may perhaps lie closer to 0.5 or 0.6.

The top panel of Figure 5 shows the posterior mean estimate of the hazard, survival and frailty density curves, as well as pointwise 95% credible bands for each. The shape of the hazard and frailty density estimates is arguably poorly described by a parametric model. We observe that the risk for readmission or death is greatest shortly after discharge from the index hospitalization, declining rapidly during the first few weeks post-discharge, then more slowly before reaching a minimum at approximately 120 days. There is a secondary peak around 150 days. This pattern of risk suggests that the design of interventions for postponing mortality or readmissions should be targeted at the care transition from the hospital to home or to another facility, with most of the benefit being realized in the first few months following the index event.

The Bayesian approach allows the marginal posterior distributions of the physician-level effects  $U_i, i = 1 \dots 31$  (i.e., frailties) to be examined. The lower

panel of Figure 5 shows boxplots that summarizes the main features of these marginal posteriors. The height and width of the boxes indicate that, with increasing cluster size, the marginal posterior distributions of the frailties exhibit less dispersion and tend to be larger than the prior mean. In the absence of heterogeneity, one would expect to see each boxplot centered approximately at the prior mean (i.e., one). The observed pattern is consistent with the existence of heterogeneity, and perhaps even suggests the possibility at least two distinct groups of physicians or defining practice characteristics.

The results summarized above rely on approximate samples from the posterior distribution obtained via MCMC. We monitor the mixing of chain parameters by examining trace plots and autocorrelation functions of the posterior samples. The trace plots for coefficients in Figure 6 indicate that the regression coefficient estimates have converged. Kernel density estimates based on the posterior samples suggest approximately normal marginal posterior distributions. However, estimates of the frailty variance do appear to mix at a lower rate than the other parameters, and show some evidence of long-term autocorrelation.

Previous studies have suggested that physician-level variables, such as physician volume and specialty, impact both the readmissions and mortality of heart failure patients (Jong *et al.*, 2003; Cujec *et al.*, 2005). The influence of the physician as the frailty effect on heart failure outcomes, as examined in our study, may be a novel contribution. However, it is also possible that the observed heterogeneity merely reflects physician-level differences not currently adjusted for in the hazard model. For example, in this data set, the physician clusters actually represent groups of physicians who share responsibilities as attending physicians for hospital care. There are inter-cluster differences in training (specialist and generalist) and also scope of practice: academic hospitalists, private practice hospitalists, and community physicians (Halasyamani *et al.*, 2005). The physician effect may also serve as a proxy for differences between in-hospital and post-hospital systems of care delivery. Community physicians are more likely to follow their patients personally after hospital discharge, whereas the academic and private hospitalists frequently transition such care to other physicians. The information available in this data set is not sufficient to examine these potential influences. Further research is needed to determine which, if any, of these effects contribute in a significant and modifiable way to heart failure outcomes.

### 5.3. Choice of model specification and priors

We now illustrate the effects of selecting an alternative model specification or different priors when analyzing the same data set. Instead of the penalized spline specification selected by the DIC criterion in Section 5.2, we fit an adaptive spline model with a parametric component, as described in Section 2.3.1.

Figure 7 shows the curves and frailties obtained using adaptive splines specified per the Adapt(20)+Par specification introduced in Section 4.2. This model specification consists of a Poisson(20) prior on the number of spline knots and models the baseline hazard and frailty density as a mixture of B-splines and a

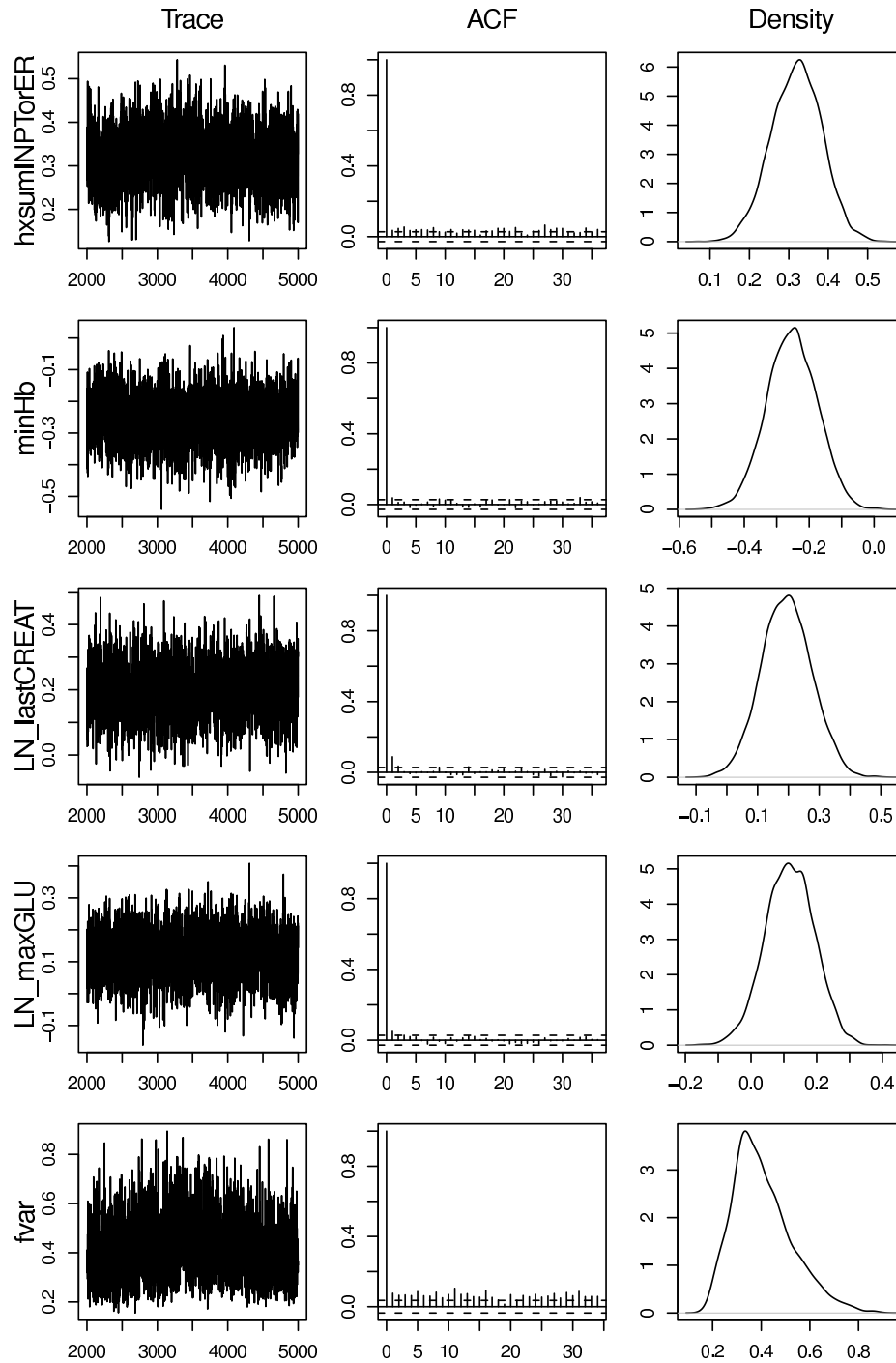


FIG 6. Trace plots, autocorrelation functions, and posterior density estimates for a regression coefficients (defined in Table 3) and frailty variance (denoted by `fvar`) of a penalized spline fit to the congestive heart failure data.

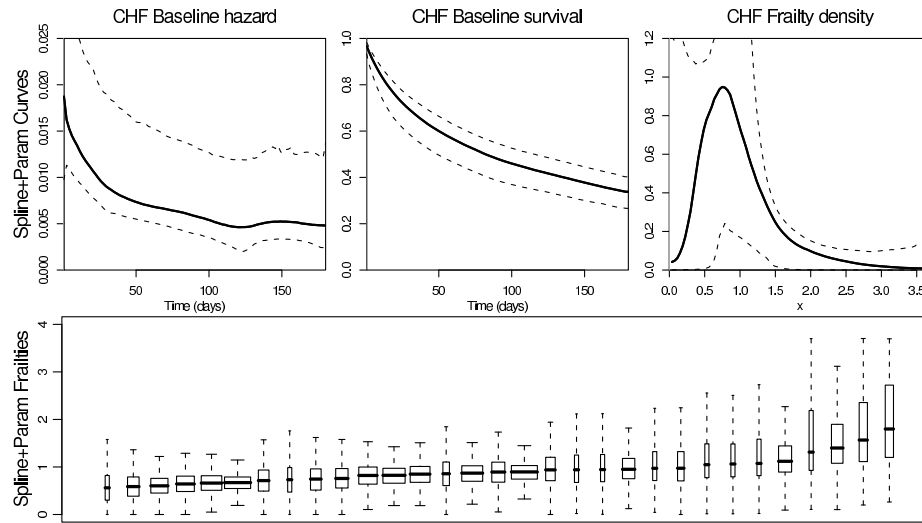


FIG 7. Upper panel: Hazard, survival and frailty density estimates, and 95% pointwise posterior intervals for the congestive heart failure data. *Adapt(20)+Parametric* model specification, with Weibull and lognormal parametric components for the hazard and frailty density respectively. Component weight prior is  $\text{Beta}(1, 2)$ , which favors the parametric component. The baseline represents a patient with average covariates. Lower panel: Boxplots of posterior frailty estimates for each of the 31 clusters, sorted in order of ascending posterior mean. Box width indicates the cluster size.

parametric curve, here being a Weibull hazard function and lognormal density, respectively. The curves in the top panel show that the specification has a substantial smoothing effect on the hazard and frailty density estimates. In particular, the parametric component dominates the frailty density, making it appear nearly lognormal, but has a smaller effect on the estimated hazard curve. The lower panel in Figure 7 shows that the marginal posterior distributions of the frailties exhibit greater variation in comparison with those in Figure 5, an effect that seems particularly noticeable with smaller clusters.

We next use the congestive heart failure data to illustrate the effect of choosing different priors on the number of knots. Figure 7 used a  $\text{Poisson}(20)$  prior on the number of knots. The top panel of Figure 8 compares this fit to those resulting from different Poisson prior choices. As expected, setting the prior to  $\text{Poisson}(1)$  leads to excessively smooth fits, whereas the  $\text{Poisson}(50)$  curve is considerably more variable and shows potentially undesirable detail. The effect on the survival curve is relatively small, however, as local bumps in the hazard are smoothed out by the integration.

The second and third panels show the effect of using geometric and negative binomial priors for the number of knots. As noted by Biller (2000), these priors universally encourage smoother fits, and are relatively insensitive to the choice of parameters. This is a desirable property if a more robust fit is preferred, but if control over the smoothness of the curve is desired, the Poisson prior is

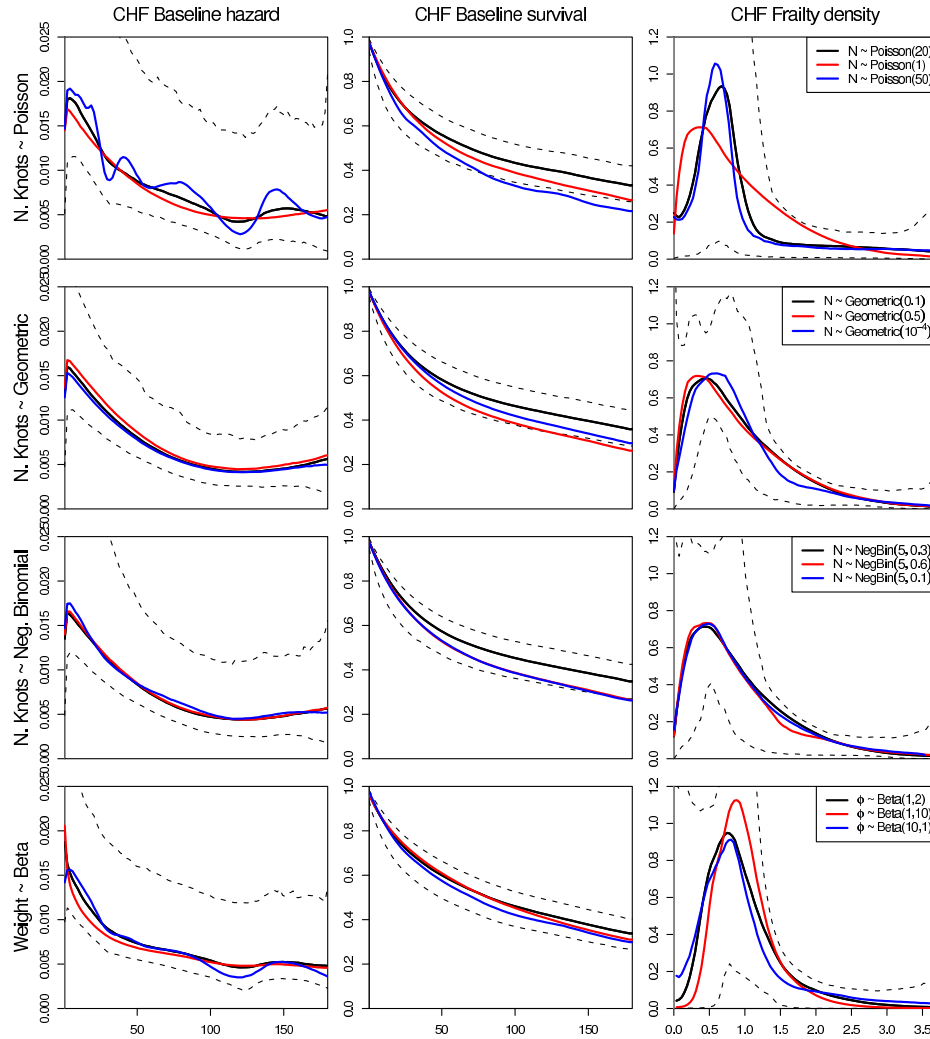


FIG 8. Illustrating the effect of choosing different priors on the number of knots, and different prior parameters for the spline weight. Credible intervals shown correspond to the black line in each plot.

preferable. DIC might be used to select the mean of the Poisson prior.

The relative prominence of the parametric and spline components can be controlled through the prior on the weights  $\phi$ . The bottom panel of Figure 8 shows the effects of changing the prior to  $\text{Beta}(1, 10)$  and  $\text{Beta}(10, 1)$ , which respectively place more and less emphasis on the parametric component. The frailty density is more sensitive to changes in the prior weight than the hazard curve, because with only 31 clusters, the data contain much less information regarding the shape of the frailty density.

## 6. Discussion

The proposed Bayesian methodology permits the analysis of clustered survival data when the underlying frailty distribution is unknown, helping to mitigate the impact associated with the specification of an incorrect parametric model on the assessment of heterogeneity. Our simulation results demonstrate that the method generally performs well, particularly so in cases where the data contain many clusters of reasonable size. From a frequentist perspective, levels of bias and also coverage (results not shown) for the posterior mean and median regression coefficients are generally comparable to those obtained under the gamma frailty model; in addition, substantial reductions in the bias of the frailty variance are frequently achieved. The results further demonstrate that unusual baseline hazards and frailty distributions can be correctly identified in the presence of sufficient information; the resulting gain in flexibility and information available from the proposed methodology may help to offset its concomitant (and potentially substantial) computational cost. Extensions of the methodology to accommodate stratification and time-dependent covariates are conceptually simple; however, at the time of this writing, such extensions are not supported by the accompanying R package.

The Bayesian approach provides a wealth of information in the form of the joint posterior distribution of all parameters of interest. MCMC sampling provides the joint posterior of the hazard and survival, as well as for cluster-level frailties; in addition, unlike standard frequentist methods of estimation for the proportional hazards frailty model, uncertainty in the frailty variance (i.e., density) is naturally reflected in the posterior behavior of all parameter estimates. Through the deliberate specification of priors, the Bayesian approach permits great flexibility in model specification. For example, with appropriate priors and hyperparameters, the smoothness of the baseline hazard and frailty density are mainly controlled through the prior distribution. In addition, in settings where either curve can be well-modeled using a parametric family of distributions, the inclusion of an appropriate parametric component results in smoother, less variable fits. Importantly, one is also not limited to prior specifications that are primarily designed to control the degree of smoothness in the baseline hazard function and frailty density; where available, practitioners can easily incorporate prior information on model parameters and the general shape of the baseline hazard function, decreasing the possibility of obscuring important effects.

The statistical methodology developed here may also prove to be a useful tool for public policy audiences. Re-hospitalizations of Medicare patients within 30 days (as a binary outcome) are receiving increasing scrutiny due to their frequency and costs (Jencks, Williams and Coleman, 2009). In addition, there is on-going dialogue regarding the current payment system and how it can be redesigned to align better with improved patient outcomes (Averill et al., 2009). Despite the interest in outcomes reported at the cluster (eg. hospital, physician) level, existing multivariable models in the clinical literature provide only a modest amount of explanatory information (Ross et al., 2008; Keenan et al., 2008), with hospital performance typically represented by normally-distributed ran-



dom intercepts. Implicit to such “profiling” approaches is that the unexplained variability may signal important quality-related performance differences at the cluster level (Keenan *et al.*, 2008). Our flexible modeling approach may help to accelerate understanding of the nature and influence of the providers in such applications, which often involve large databases, through modeling provider-level effects in a more nonparametric way. Understanding the provider characteristics associated with the best (and worst) outcomes in such a distribution may help in identifying the best practices to replicate.

## 7. Acknowledgements

We would like to thank our two referees for their helpful and constructive comments, and also our “guest editor” Larry Wasserman for his assistance in handling this paper. Emmanuel Sharef and Robert Strawderman gratefully acknowledge the support of NIH grant GM056182; David Ruppert gratefully acknowledges the support of NIH grant CA57030 and NSF grant DMS-0805975.

## Appendix A: Choice of penalty functions

### A.1. Gaussian penalty

The penalty function may be chosen to yield a Gaussian prior on the parameter set, for example  $p_\lambda(\boldsymbol{\theta}_\lambda) = \boldsymbol{\theta}_\lambda^T \boldsymbol{\theta}_\lambda$ , and analogously for  $p_u(\boldsymbol{\theta}_u)$ . This penalty function is recommended when adaptive knot selection is used, since in that case, the smoothness of the curve is controlled through the prior on the number of knots, and does not need to be explicitly penalized.

### A.2. Penalty on second differences

Let  $\mathbf{D}$  be a matrix so that  $\mathbf{D}\mathbf{y}$  computes the second difference in  $\mathbf{y}$ , and let  $\mathbf{P} = \mathbf{D}^T \mathbf{D}$ . Then, for analogously defined matrices  $\mathbf{P}_\lambda, \mathbf{P}_u$  of the appropriate dimensions, the following functions penalize the second differences in the spline parameters:

$$p_\lambda(\boldsymbol{\theta}_\lambda) = \boldsymbol{\theta}_\lambda^T \mathbf{P}_\lambda \boldsymbol{\theta}_\lambda, \quad p_u(\boldsymbol{\theta}_u) = \boldsymbol{\theta}_u^T \mathbf{P}_u \boldsymbol{\theta}_u.$$

While this choice of penalty function is appropriate when the knots are equally spaced, it does not necessarily result in smooth behavior otherwise. When knots are unevenly spaced, it is possible to adjust the matrix  $\mathbf{D}$  to compute appropriately scaled differences.

### A.3. Penalty on the second derivative

In order to ensure smoothness even when knots are not equally spaced, we can construct a penalty on the second derivative of the spline. In the case of the

baseline hazard spline, that is

$$p_\lambda(\boldsymbol{\theta}_\lambda) = \int_0^\infty \left( \lambda_0^{(2)}(t, \boldsymbol{\theta}_\lambda) \right)^2 dt = \int_0^\infty \left( \sum_{k=1}^{K_\lambda} B_{\lambda k}^{(2)}(t) \exp(\theta_{\lambda k}) \right)^2 dt = e^{\boldsymbol{\theta}_\lambda^T \mathbf{P}_\lambda} e^{\boldsymbol{\theta}_\lambda},$$

where  $\mathbf{P}_\lambda$  is a matrix whose  $(j, k)$  entry is

$$P_{\lambda, jk} = \int_0^\infty B_{\lambda j}^{(2)}(t) B_{\lambda k}^{(2)}(t) dt.$$

We can construct an analogous penalty matrix for the frailty density, keeping in mind that the frailty density uses normalized splines, that is,

$$p_u(\boldsymbol{\theta}_u) = \frac{e^{\boldsymbol{\theta}_u^T \tilde{\mathbf{P}}_u} e^{\boldsymbol{\theta}_u}}{(\mathbf{1}_{K_u}^T e^{\boldsymbol{\theta}_u})^2}, \quad \text{where} \quad \tilde{P}_{u, jk} = \int_0^\infty \tilde{B}_{uj}^{(2)}(t) \tilde{B}_{uk}^{(2)}(t) dt.$$

## Appendix B: Choice of parametric components

Both the baseline hazard and frailty density may have optional parametric components. In this section, we present some of the possible choices of distributions, parametrized in a way that allows Gaussian priors:

An exponential baseline hazard can be parametrized by a constant log-hazard  $\boldsymbol{\eta}_\lambda = \eta_\lambda$ , so that the hazard function is

$$\lambda_{0p}(t, \boldsymbol{\eta}_\lambda) = \exp(\eta_\lambda), \quad \Lambda_{0p}(t, \boldsymbol{\eta}_\lambda) = t \exp(\eta_\lambda).$$

Similarly, the Weibull baseline hazard is parametrized by a log-hazard  $\eta_{\lambda 1}$  and log scale parameter  $\eta_{\lambda 2}$ , so the hazard function is

$$\lambda_{0p}(t, \boldsymbol{\eta}_\lambda) = \exp(\eta_{\lambda 1} + \eta_{\lambda 2} + (e^{\eta_{\lambda 2}} - 1) \log t), \quad \Lambda_{0p}(t, \boldsymbol{\eta}_\lambda) = \exp(\eta_{\lambda 1}) t^{\exp(\eta_{\lambda 2})}.$$

For the frailty distribution, we parametrize a gamma frailty distribution by its log-variance  $\eta_u$ , resulting in the following parametric density component:

$$f_p(x, \eta_u) = x^{\exp(-\eta_u) - 1} \frac{\exp(-\eta_u) \exp(-\eta_u) e^{-\exp(-\eta_u)x}}{\Gamma(\exp(-\eta_u))}.$$

Similarly, the lognormal frailty density parametrized by a log-variance parameter  $\eta_u$  gives:

$$f_p(x, \eta_u) = \frac{\exp\left(-\frac{(\log x + \frac{1}{2} \exp(\eta_u))^2}{2 \exp(\eta_u)}\right)}{x \sqrt{2\pi \exp(\eta_u)}}.$$

## Supplementary Material

**Supplementary materials to accompany “Bayesian adaptive B-spline estimation in proportional hazards frailty models” by Sharef, Strawderman, Ruppert, Cowen, and Halasyamani**

(doi: [10.1214/10-EJS566SUPP](https://doi.org/10.1214/10-EJS566SUPP)). The supplementary materials contain an expanded version of Section 3, including the various conditional posteriors used in our Gibbs sampling algorithm and adaptive knot selection stages.

## References

- ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. and KEIDING, N. (1993). *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York. [MR1198884](#)
- ASLANIDOU, H., DEY, D. K. and SINHA, D. (1998). Bayesian Analysis of Multivariate Survival Data Using Monte Carlo Methods. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* **26** 33–48. [MR1624365](#)
- AVERILL, R. F., MCCULLOUGH, E. C., HUGHES, J. S., GOLDFIELD, N. I., VERTREES, J. C. and FULLER, R. L. (2009). Redesigning the Medicare inpatient PPS to reduce payments to hospitals with high readmission rates. *Health Care Financ Rev* **30** 1–15.
- BARKER, P. and HENDERSON, R. (2005). Small Sample Bias in the Gamma Frailty Model for Univariate Survival. *Lifetime Data Analysis* **11**. [MR2158785](#)
- BILLER, C. (2000). Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models. *Journal of Computational and Graphical Statistics* **9** 122–140. [MR1819868](#)
- CLAYTON, D. G. (1991). A Monte Carlo Method for Bayesian Inference in Frailty Models. *Biometrics* **47** 467–485.
- CLAYTON, D. and CUZICK, J. (1985). Multivariate Generalizations of the Proportional Hazards Model. *Journal of the Royal Statistical Society. Series A* **148** 82–117. [MR0806480](#)
- COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. [MR0341758](#)
- COX, D. R. (1975). Partial Likelihood. *Biometrika* **62** 269–276. [MR0400509](#)
- CUJEC, B., QUAN, H., JIN, Y. and JOHNSON, D. (2005). Association between physician specialty and volumes of treated patients and mortality among patients hospitalized for newly diagnosed heart failure. *Am J Med* **118** 35–44.
- DE BOOR, C. (2001). *A Practical Guide to Splines*. Applied Mathematical Sciences **27**. Springer. [MR1900298](#)
- DENISON, D. G. T., MALLICK, B. K. and SMITH, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60** 333–350. [MR1616029](#)
- DIMATTEO, I., GENOVESE, C. R. and KASS, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88** 1055–1071. [MR1872219](#)
- FISHER, L. D., DIXON, D. O., HERSON, J., FRANKOWSKI, R. F., HEARON, M. S. and PEACE, K. E. (1990). Intention to Treat in Clinical Trials. In *Statistical Issues in Drug Research and Development (Karl E. Peace, Editor)* 331–350. Marcel Dekker, New York.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC. [MR2027492](#)
- GHIDEY, W., LESAFFRE, E. and EILERS, P. (2004). Smooth Random Effects Distribution in a Linear Mixed Model. *Biometrics* **60** 945–953. [MR2133547](#)
- GHOSH, J. K., DELAMPADY, M. and SAMANTA, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods*. Springer-Verlag. [MR2247439](#)

- GREEN, P. J. (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika* **82** 711–732. [MR1380810](#)
- HALASYAMANI, L. K., VALENSTEIN, P. N., FRIEDLANDER, M. P. and COWEN, M. E. (2005). A comparison of two hospitalist models with traditional care in a community teaching hospital. *Am J Med* **118** 536–543.
- HJORT, N. L. and GLAD, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist.* **23** 882–904. [MR1345205 \(96j:62083\)](#)
- HJORT, N. L. and JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24** 1619–1647. [MR1416653 \(97m:62034\)](#)
- IBRAHIM, J. G., CHEN, M.-H. and SINHA, D. (2001). *Bayesian survival analysis. Springer Series in Statistics*. Springer-Verlag, New York. [MR1876598 \(2002j:62002\)](#)
- JENCKS, S. F., WILLIAMS, M. V. and COLEMAN, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *N Engl J Med* **360** 1418–1428.
- JONG, P., GONG, Y., LIU, P. P. et al. (2003). Care and outcomes of patients newly hospitalized for heart failure in the community treated by cardiologists compared with other specialists. *Circulation* **108** 184–191.
- KEENAN, P. S., NORMAND, S.-L. T., LIN, Z., DRYE, E. E., BHAT, K. R., ROSS, J. S., SCHUUR, J. D., STAUFFER, B. D., BERNHEIM, S. M., EPSTEIN, A. J., WANG, Y., HERRIN, J., CHEN, J., FEDERER, J. J., MATTERA, J. A., WANG, Y. and KRUMHOLZ, H. M. (2008). An administrative claims measure suitable for profiling hospital performance on the basis of 30-day all-cause readmission rates among patients with heart failure. *Circ Cardiovasc Qual Outcomes* **1** 29–37.
- KOMÁREK, A. (2006). Accelerated Failure Time Models for Multivariate Interval-Censored Data with Flexible Distributional Assumptions PhD thesis, Katholieke Universiteit Leuven.
- KOMÁREK, A. and LESAFFRE, E. (2008). Bayesian Accelerated Failure Time Model with Multivariate Doubly-Interval-Censored Data and Flexible Distributional Assumptions. *Journal of the American Statistical Association* **103** 523–533.
- KOOPERBERG, C., STONE, C. J. and TRUONG, Y. K. (1995). Hazard Regression. *Journal of the American Statistical Association* **90** 78–94. [MR1325116](#)
- KOSOROK, M. R., LEE, B. L. and FINE, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Ann. Statist.* **32** 1448–1491. [MR2089130 \(2005m:62181\)](#)
- LEBLANC, M. and CROWLEY, J. (1999). Adaptive Regression Splines in the Cox Model. *Biometrics* **55** 204–213.
- LINDSTROM, M. J. (2002). Bayesian estimation of free-knot splines using reversible jumps. *Computational Statistics & Data Analysis* **41** 255–269. [MR1945871](#)
- MALLICK, B. K., DENISON, D. G. T. and SMITH, A. F. M. (1999). Bayesian Survival Analysis Using a MARS Model. *Biometrics*, **55** 1071–1077.
- MÜLLER, P. and QUINTANA, F. A. (2004). Nonparametric Bayesian Data Analysis. *Statist. Sci.* **19** 95–110. [MR2082149](#)

- NASKAR, M. (2008). Semiparametric analysis of clustered survival data under nonparametric frailty. *Stat Neer* **62** 155–172. [MR2427867](#)
- NIELSEN, G. G., SØRENSEN, T. I. A., GILL, R. D. and ANDERSEN, P. K. (1992). A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics* **19** 25–43. [MR1172965](#)
- PAN, W. (2001). Using Frailties in the Accelerated Failure Time Model. *Lifetime Data Analysis* **7** 55–64. [MR1836201](#)
- ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*. Springer Series in Statistics. Springer, New York. [MR2080278](#)
- ROSS, J. S., MULVEY, G. K., STAUFFER, B., PATLOLLA, V., BERNHEIM, S. M., KEENAN, P. S. and KRUMHOLZ, H. M. (2008). Statistical models and patient predictors of readmission for heart failure: a systematic review. *Arch Intern Med* **168** 1371–1386.
- RUPPERT, D., NETTLETON, D. and HWANG, J. T. G. (2007). Exploring the Information in p-Values for the Analysis and Planning of Multiple-Test Experiments. *Biometrics* **63** 483–495. [MR2370807](#)
- SHAREF, E. (2008). Nonparametric Frailty Models for Clustered Survival Data PhD thesis, Cornell University.
- SHAREF, E., STRAWDERMAN, R. L., RUPPERT, D., COWEN, M. and HALASYAMANI, L. (2010). Supplement to “Bayesian Adaptive B-spline Estimation in Proportional Hazards Frailty Models.” DOI: [10.1214/10-EJS566SUPP](#).
- SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75** 317–343.
- SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 583–639. [MR1979380](#)
- STAUDENMAYER, J., RUPPERT, D. and BUONACCORSI, J. P. (2008). Density estimation in the presence of heteroskedastic measurement error. *Journal of the American Statistical Association* **103** 726–736.
- THERNEAU, T. M. and GRAMBSCH, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer. [MR1774977](#)
- WALKER, S. G. and MALLICK, B. K. (1997). Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing. *Journal of the Royal Statistical Society. Series B (Methodological)* **59** 845–860. [MR1483219](#)