Bayesian variable selection in the AFT model with an application to the SEER breast cancer data

Zhen Zhang[1], Samiran Sinha[2,*], Tapabrata Maiti[3], and Eva Shipp[4]
[1] Department of Statistics, University of Chicago, Chicago, Illinois
[2] Department of Statistics, Texas A&M University, College Station, Texas
[3] Department of Statistics and Probability, Michigan State University, East Lansing, Michigan
[4] Texas A&M Health Science Center, School of Rural Public Health, College Station, Texas
[*] email: sinha@stat.tamu.edu

Summary

Accelerated failure time (AFT) model is a popular model to analyze censored time-to-event data. Analysis of this model without assuming any parametric distribution for the model error is challenging, and the model complexity is enhanced in the presence of large number of covariates. We develop a noparametric Bayesian method for regularized estimation of the regression parameters in a flexible AFT model. The novelties of our method lie in modeling the error distribution of the accelerated failure time non-parametrically, modelling the variance as a function of the mean, and adopting a variable selection technique in modeling the mean. The proposed method allowed for identifying a set of important regression parameters, estimating survival probabilities, and constructing credible intervals of the survival probabilities. We evaluated operating characteristics of the proposed method via simulation studies. Finally, we apply our new comprehensive method to analyze the motivating breast cancer data from the Surveillance, Epidemiology, and End Results (SEER) Program, and estimate the 5-year survival probabilities for women included in the SEER database who were diagnosed with breast cancer between 1990 and 2000.

Key Words: Accelerated failure time; Bayesian Lasso; Dirichlet process prior; Markov chain Monte Carlo; Prognostic factors; Survival probability.

Running title: Variable selection in an AFT model

1 Introduction

We start this section with some discussions of the motivating data. The Surveillance, Epidemiology, and End Results (SEER) Program [1] routinely collects population-based cancer patient data from 20 registries across the United States. In existence since 1973 and housed within the National Cancer Institute (NCI), it is one of the most important, comprehensive, and widely used sources of information for studying the survival of cancer patients. It is of scientific and public interest to estimate 5-year survival probabilities for cancer patients and to understand how prognostic and demographic factors impact survival. This information is routinely used in determining treatment plans and for making policy decisions. Importantly, the National Cancer Institute uses these survival probabilities to compute the 5-year relative survival rates (`http://seer.cancer.gov/archive/publications/survival/`).

It is well-known that overall these survival probabilities depend, at least in some degree, on the following 8 disease characteristics or prognostic factors: 1) stage of the disease at the time of diagnosis, 2) tumor grade, 3) histology, 4) tumor size, 5) extension of the primary tumor, 6) nodal involvement, 7) estrogen receptor (ER) status, and 8) progesterone receptor (PR) status, and two demographic characteristics: 9) age at diagnosis and 10) race. As documented in a SEER publication [2, Chap. 13], these characteristics not only exert main effects on the relative survival rate but there also exists an interaction effect of any two characteristics. The survival probabilities were calculated based on the life-table approach [2, p. 102] considering at most two characteristics at a time. In some occasions, due to small sample sizes, the survival probabilities were not calculated, resulting in some empty cells in Tables 13.5, 13.6 and 13.8 of the document, for example. Also of critical concern, the published rates do not accompany any uncertainty measures such as the standard error.

To calculate more meaningful survival probabilities, one thus needs to take into account several factors (not just one or two) and their interactions simultaneously. Motivated by this scenario, we consider an accelerated failure time (AFT) model and propose a regularized estimation method.

Due to having the linear model structure and easy interpretation of the model parameters, accelerated failure time (AFT) model is a popular choice after the proportional hazard model for analyzing censored data. Suppose T is the time-to-event and $\boldsymbol{Z}$ is a q-vector of covariates, then under the AFT model

$$Y = \log(T) = \boldsymbol{Z}^T\boldsymbol{\beta} + e,$$

where $\boldsymbol{\beta}$ denotes the regression parameter for $\boldsymbol{Z}$, and e denotes the residual term. A parametric AFT model is obtained if a parametric distribution is adopted for the residual e, whereas the nonparametric AFT model is obtained when the distribution of e is left unspecified with some mild regularity conditions. The nonparametric AFT model is studied extensively in the frequentist arena ([3]; [4]; among others) as well as in the Bayesian paradigm ([5]; [6]; [7]). To date, most papers reporting on the use of the AFT model include, as an assumption, that the residual term of the model is independent of the predictors.

Here we are concerned about the selection of important variables and estimation of the regression coefficients when q is large, and in this case regularized estimators are commonly used where estimators are obtained by maximizing a penalized objective function. In that respect least shrinkage selection operator (LASSO) has been widely used. If there is no censored observation then the LASSO estimators are obtained by minimizing

$$\sum_{i=1}^{n}\{Y_i - \overline{Y} - (\boldsymbol{Z}_i - \overline{\boldsymbol{Z}})^T\boldsymbol{\beta}\}^2 + \lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ and $\overline{\boldsymbol{Z}} = \sum_{i=1}^{n}\boldsymbol{Z}_i/n$, and $\lambda > 0$ is the penalty parameter. In the classical paradigm, the unknown penalty parameter is determined by some cross-validation method. Tibshirani [8] first used the LASSO method for variable selection in the Cox regression model. Although LASSO works well for the best subset selection, generally the non-zero parameters are estimated with asymptotic bias ([9], [10]). Zou [11] showed that in a linear regression model, under some nontrivial conditions, LASSO satisfies oracle properties (identifies true non-zero set of covariates and for the non-zero coefficients, the asymptotic distribution of the estimator minus the true parameter follows a mean-zero normal distribution). Additionally, Huang et al. [12] investigated the oracle property of the LASSO estimator in the Cox model for sparse and high-dimensional covariates (i.e., $q >> n$).

In the nonparametric AFT model, Huang et al. [13] showed that the LASSO is asymptotically consistent when q is fixed, and n gets large. Their result even covers the case where the variance of e may depend on the covariates. Although there are several works on Bayesian variable selection in the parametric AFT model ([14]; [15]), as far as we know, there is no Bayesian variable selection work in the nonparametric AFT model. The aim of this paper is to apply the Bayesian variable selection technique using LASSO in the AFT model where the residual e is modeled nonparametrically. There are two reasons for considering the AFT model, 1) easy interpretation of the model parameters and 2) the Bayesian variable selection has been largely unexplored in a nonparametric setting.

One big advantage of the Bayesian regularization is that the penalty parameter can be easily estimated by putting a prior on this. Then the Bayesian mechanism allows the prior to integrate with the data. Secondly, we treat the residual e nonparametrically by modeling its distribution via a Dirichlet mixture of normal densities. We use a Dirichlet process mixture of normals to gain complete flexibility of the model. It is known that a mixture of normal distributions is more flexible than a single normal distribution that requires one to specify the number of mixing components. On the other hand, in a Dirichlet process mixture of normals, that theoretically allows an infinite number of mixing components, the number of mixing components is not fixed but allowed to be determined in a data-driven way resulting in a more flexible model for e. Finally, we relax the independence assumption between the residual and the predictors– what this means is that different groups formed by a level combination of the predictor variables, not only have different means for the survival time, but may also have different variances. It is well known that proper modeling of the variance not only increases accuracy of the regression coefficients, but in our case, will accurately measure the survival probability. This comes at the cost of increased complexity compared to the case where the residual term of the AFT model and covariates are assumed independent. We assume that the variance is a polynomial function of the mean. In the generalized estimating equation context, Chiou and Müller [16] modeled the variance as a smooth function of the mean. Kauermann and Wegener [17] discussed variance modeling in some other contexts. But to date, no one has considered variance modeling in the context of an AFT model.

A brief outline of the rest of the article is as follows. Section 2 introduces models, assumptions, and priors. Posterior computation is given in Section 3. Section 4 provides an estimation of the survival probabilities. Sections 5 and 6 then contain the simulation study and analysis of the SEER breast cancer data, respectively. Conclusions are given in Section 7.

2 Models, assumptions, and priors

Mimicking the real data, suppose that the observed data are $(V_i, \Delta_i, X_{i1}, \ldots, X_{ip})$, $i = 1, \ldots, n$. Here $V_i = \min(T_i, C_i)$, the minimum of the survival time T_i and the random censoring time C_i whichever occurred earlier, for the ith subject. Also, Δ_i denotes the censoring indicator $\Delta_i = I(T_i \leq C_i)$. Assume that T and C are independent conditional on the given covariates. For our breast cancer data C is either the end of the follow-up time which is December 31, 2003 (SEER 1973-2003 Public-Use CD) or the last time SEER had information about the subject, and $X_1, \cdots, X_p$ are p factors that include prognostic and demographic factors.

Model for the residual term: We consider the following linear model

$$
\begin{aligned}
\log(T_i) &= \mu_i + e_i, \\
\mu_i &= \boldsymbol{Z}_i^T \boldsymbol{\beta}, \\
e_i &= \sqrt{v(\mu_i)}\varepsilon_i, \\
\varepsilon_i | \theta_i = (\theta_{i1}, \theta_{i2}) &\sim \text{Normal}(\theta_{i1}, (\sqrt{\theta_{i2}})^2), \\
\theta_i | \mathscr{P} &\sim \mathscr{P}, \text{ a random probability measure,}
\end{aligned}
$$

$$\mathscr{P} \sim \mathcal{P}_N,$$

where $\boldsymbol{\beta}$ is the regression parameter corresponding to $\boldsymbol{Z}$ which comprises of all the main effects and two factor interactions of $(X_1, \cdots, X_p)$, and $v(\mu)$ is a positive valued function. Note that $\mathscr{P}$ is a random probability measure on $(\mathscr{R} \times \mathscr{R}^+, \mathscr{B})$, where $\mathscr{B}$ denotes the Borel σ-algebra defined on $\mathscr{R} \times \mathscr{R}^+$. The stick-breaking prior $\mathcal{P}_N$, identified by the precision parameter α and the base probability measure $H(\cdot|\psi)$, is almost surely a discrete random probability measure. Now

$$\mathcal{P}_N(\cdot) = \sum_{k=1}^{N} p_k \delta_{\vartheta_k}(\cdot),$$

where ϑ_k are i.i.d. random variables from a base probability measure H (the corresponding density is denoted by h), and p_ks are random variables chosen to be independent of ϑ_k such that $0 \leq p_k \leq 1$ and $\sum_{k=1}^{N} p_k = 1$, and $(p_1, \cdots, p_N) \sim \text{Dirichlet}(\alpha/N, \ldots, \alpha/N)$. When $N \to \infty$ the stick-breaking prior becomes the well-known Dirichlet process prior, usually written as $DP(\alpha H)$. Therefore, sometime $\mathcal{P}_N$ is referred to as a finite dimensional Dirichlet process. Here α plays a critical role in the variance of the random probability measure. On the precision parameter α we shall use a $\text{Gamma}(a_\alpha, b_\alpha)$ prior. Importantly, a Dirichlet process mixture of normal distributions covers a large class of densities with finite variance, and the corresponding posterior distribution is weakly consistent for the true density ([18], p. 152).

We would like to point out that Christensen and Johnson [5] used a Dirichlet process prior directly on the residual term of the linear model of $\log(T)$, and proposed a semi-Bayes approach to make inference on the regression parameters. On the other hand, we not only deal with the regression parameters associated with the mean function, but also our variance function varies with the mean function making it difficult to adopt Christensen and Johnson's approach in our set-up.

Model for the variance: Suppose that the conditional variance of $\log(T)$ is $v(\mu)$, where v is assumed to be a function of μ known up to a finite dimensional parameter. We shall model v using a polynomial function of μ, such as $v_i = v_i(\mu_i) = \exp(\sum_{l=1}^{L} \gamma_l \mu_i^l)$, for some L. In particular, $v_i = 1$

when $\mu_i = 0$. Here L is assumed to be fixed, and one may apply a model selection technique to choose an optimal value of L. To facilitate Bayesian computation we use a hierarchical structure,

$$\begin{aligned} v_i = \mathrm{var}\{\log(T_i)|Z_i\} &= \exp(\gamma_1\eta_i + \gamma_2\eta_i^2 + \cdots + \gamma_L\eta_i^L), \\ \eta_i &= \mu_i + e_{i\eta}, \\ e_{i\eta} &\sim \mathrm{Normal}(0, \tau_\eta^2). \end{aligned}$$

This hierarchical structure helps easy drawing of the β-parameters in the Gibbs sampling. Usually τ_η^2 is chosen to be a very small number. On the unknown parameters $\gamma_1, \ldots, \gamma_L$ we shall use independent Normal$(m_{\gamma_l}, \sigma_{\gamma_l}^2)$, $l = 1, \ldots, L$ priors.

Handling large dimension of $\boldsymbol{Z}$*:* A large number of parameters may increase predictability of a model at the cost of large uncertainty. Therefore, to reduce the effective dimension of our model we shall adopt the idea of penalized regression which will lead to a small group of variables with good prediction accuracy. In the Bayesian context inferences are based on a model averaging technique. Although there are numerous approaches to penalize the regression parameters, we shall adopt the Least Absolute Shrinkage and Selection Operator (LASSO) proposed by Tibshirani [19]. In the Bayesian context the LASSO estimator is obtained from the linear regression with the Laplace prior $\pi(\beta_1, \cdots, \beta_q) = \prod_{j=1}^q (\lambda/2)\exp(-\lambda|\beta_j|)$ on the regression coefficients. Thus the prior for $\beta_1, \cdots, \beta_q$ will be

$$\pi(\beta_1, \cdots, \beta_q) = \prod_{j=1}^q \frac{\lambda}{2}\exp(-\lambda|\beta_j|),$$

and the parameter inference will be based on the censored data likelihood and the priors. To facilitate the computation, following Park and Casella [20] we shall write the Laplace prior as a gamma mixture of a normal distribution

$$\pi(\beta_1, \cdots, \beta_q) = \prod_{j=1}^q \frac{\lambda}{2}\exp\Big(-\lambda|\beta_j|\Big) = \prod_{j=1}^q \int_0^\infty \frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\Big(-\frac{\beta_j^2}{2\sigma_j^2}\Big)\frac{\lambda^2}{2}\exp(-\lambda^2\sigma_j^2/2)d\sigma_j^2.$$

On the lasso parameter λ^2 (not on λ) we put the following Gamma(r, δ) prior

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}\exp(-\delta\lambda^2),\ r > 0,\ \delta > 0. \tag{1}$$

3 Posterior computation

When a subject is censored, we assume its actual time-to-event is T_i^* that is an unobserved latent variable and $T_i^* > V_i$. Define $\boldsymbol{T}^* = \{T_i^* : \Delta_i = 0\}$, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^T$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^T$, $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_N^*)^T$. Then the joint posterior distribution of all the parameters and the latent variables is

$$\begin{aligned}
&\pi(\boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_q^2, \lambda^2, \boldsymbol{\gamma}, p_1, \ldots, p_N, \boldsymbol{\theta}^*, \boldsymbol{\eta}, \alpha, \boldsymbol{T}^*|\text{Data})\\
\propto\ &\prod_{i=1}^{n}\Bigg(\sum_{k=1}^{N} p_k\delta_{\vartheta_k}(\theta_i)\exp\Bigg[-\frac{\Delta_i}{2\theta_{i2}}\Bigg\{\frac{\log(V_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{\sqrt{\exp(\sum_{l=1}^{L}\gamma_l\eta_i^l)}} - \theta_{i1}\Bigg\}^2\\
&-\frac{(1-\Delta_i)I(T_i^* > V_i)}{2\theta_{i2}}\Bigg\{\frac{\log(T_i^*) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{\sqrt{\exp(\sum_{l=1}^{L}\gamma_l\eta_i^l)}} - \theta_{i1}\Bigg\}^2\Bigg]\\
&\times\frac{1}{\theta_{i2}^{1/2}} \times \exp\Bigg\{-\frac{1}{2}\sum_{l=1}^{r}\gamma_l\eta_i^l - \frac{(\eta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau_\eta^2} - \frac{1}{2}\log(\tau_\eta^2)\Bigg\}\Bigg)\\
&\times\prod_{j=1}^{q}\frac{\lambda^2}{2\sqrt{2\pi\sigma_j^2}}\exp\Bigg(-\frac{\beta_j^2}{2\sigma_j^2} - \frac{\lambda^2\sigma_j^2}{2}\Bigg)\times\frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}\exp(-\delta\lambda^2)\\
&\times\prod_{l=1}^{L}\frac{1}{\sqrt{\sigma_{\gamma_l}^2}}\exp\Bigg\{-\frac{(\gamma_l - m_{\gamma_l})^2}{2\sigma_{\gamma_l}^2}\Bigg\}\\
&\times\frac{\Gamma(\alpha)}{\{\Gamma(\alpha/N)\}^N}p_1^{\alpha/N-1} \times \cdots \times p_N^{\alpha/N-1} \times \alpha^{a_\alpha-1}\exp\left(-\frac{\alpha}{b_\alpha}\right) \times \prod_{j=1}^{N} h(\vartheta_j|\psi).
\end{aligned}$$

We shall estimate the parameters via the Gibbs sampling algorithm, where we repeatedly sample the unknown parameters from their full conditional distributions. In particular, the following 10 steps will be repeated for 20,000 times for our simulation study and in the data example. Define $\boldsymbol{Z}_{i(-j)} = (Z_{i1}, \cdots, Z_{i(j-1)}, Z_{i(j+1)}, \cdots, Z_{iq})$, $\boldsymbol{\beta}_{(-j)} = (\beta_1, \cdots, \beta_{(j-1)}, \cdots, \beta_{(j+1)}, \cdots, \beta_q)$. Define $T_i^* =$

V_i for $\Delta_i = 1$ and when $\Delta_i = 0$ initialize T_i^* by some number larger than V_i. Define $\phi(a, b, c) = \exp\{-(a-b)^2/2c\}/\sqrt{2\pi c}$ and $\Phi(a) = \int_{-\infty}^{a} \phi(u, 0, 1)du$. For handling the stick-breaking prior, we introduce the configuration indicators $\boldsymbol{s} = (s_1, \cdots, s_n)$ that are defined as follows: $s_i = j$ if $\theta_i = \vartheta_j$ for $1 \leq j \leq N$ for $i = 1, \cdots, n$. Also, we define the cluster size m_j as the number of s_is equal to j. Thus, $0 \leq m_j \leq n$ and $\sum_{j=1}^{N} m_j = n$. For the base probability measure H of $\vartheta_j = (\vartheta_{j1}, \vartheta_{j2})$, we assume that $[\vartheta_{j1}|\vartheta_{j2}] \sim \text{Normal}(\psi_1, \zeta\vartheta_{j2})$, $\vartheta_{j2} \sim \text{IG}(\psi_2, \psi_3)$, and $\zeta \sim \text{IG}(\psi_4, \psi_5)$. Here we shall assume that the hyperparameters $\psi = (\psi_1, \ldots, \psi_5)^T$, m_{γ_l}, $\sigma^2_{\gamma_l}$, $l = 1, \ldots, L$, δ, r, a_α and b_α will be specified by practitioners. We shall use $\boldsymbol{Z}$ to denote the $n \times q$ design matrix $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^T$. Before we start the MCMC iterations, we initialize $\beta_1, \ldots, \beta_q, \sigma_1^2, \cdots, \sigma_q^2$, λ^2, γ_l, $l = 1, \ldots, L$, $p_1, \cdots, p_N$, $\vartheta_1, \ldots, \vartheta_N$, $e_{i\eta}$, $i = 1, \cdots, n$, ψ, α. Also, we initialize $\boldsymbol{s}$ by generating random numbers from the discrete uniform$(1, N)$, and accordingly calculate m_j, $j = 1, \ldots, N$. The detailed MCMC steps are given in the Appendix.

In Steps 4 and 6 given in the Appendix for updating $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ that require Metropolis step, we choose the normal proposal with diminishing proposal variance based on the adaptive MCMC techniques proposed in [21]. This adaptive MCMC helps to achieve a reasonable and recommended acceptance rate, say $\pi_o = 44\%$, that is required for good mixing. More specifically, suppose that for a generic variable x, we use normal proposal density with the current value of the variable as the mean and κ as the variance. In the Markov chain, we treat every $B = 50$ iterations as a batch. For the bth batch with B iterations, we check the acceptance rate $\pi^b(x)$ for a generic variable x within that batch. If $\pi^b(x)$ is greater (smaller) than π_o, we then subtract (add) $\xi = \min\{0.01, b^{-1/2}\}$ from (to) the logarithm of the proposal standard deviation $\log(\kappa^{1/2})$. It is clear that the small tuning amount ξ diminishes as the batch number b increases. Finally, based on the MCMC samples we compute the posterior mean, 95% credible interval for each of the β-parameter, and most importantly estimate the survival probabilities.

4 Estimation of survival probabilities

Observe that conditional on $\boldsymbol{Z} = \boldsymbol{Z}_0$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\theta = (\theta_1, \theta_2)^T$,

$$\mathrm{pr}(T > t_0 | \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \theta) = 1 - \Phi\bigg(\frac{\log(t_0) - \mu_0 - \sqrt{v(\mu_0, \gamma)}\theta_1}{\sqrt{v(\mu_0, \gamma)\theta_2}} \bigg),$$

where $v(\mu_0, \gamma) = \exp(\sum_{l=1}^{L} \gamma_l \mu_0^l)$ with $\mu_0 = \boldsymbol{Z}_0^T \boldsymbol{\beta}$. Thus, the posterior distribution of this survival probability obtained by recording $(1 - \sum_{k=1}^{N} p_{jk} \Phi[\{\log(t_0) - \boldsymbol{Z}_0^T \boldsymbol{\beta}_j - \sqrt{v(\boldsymbol{Z}^T \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j)}\vartheta_{k1}^{(j)}\}\{v(\boldsymbol{Z}^T \boldsymbol{\beta}_j, \boldsymbol{\gamma}_j)$ $\vartheta_{k2}^{(j)}\}^{-1/2}])$ for $j = 1, \cdots, M$, where $\boldsymbol{\beta}_j, \boldsymbol{\gamma}_j, p_{jk}, \vartheta_k^{(j)} = (\vartheta_{k1}^{(j)}, \vartheta_{k2}^{(j)})^T$, $k = 1, \cdots, N$ are the M MCMC samples from the posterior distribution of the parameters. In particular, note that p_{jk} and $\vartheta_k^{(j)}$ are coming from Steps 8 and 9 given in the Appendix. The average of these M probabilities is the estimated posterior mean.

5 Simulation study

Simulation design: In this simulation we assess the small sample performance of our method and compare it with other approaches. We simulated cohort data with $n = 5,000$ by simulating $Z_1, \cdots, Z_{20}$ independently from the Bernoulli(0.5) distribution. However, the mean involves only Z_1, Z_2, Z_3, and Z_4, and write $\mu = \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4$ with $\beta_1 = \beta_2 = 0.5$, $\beta_3 = 0.35$, and $\beta_4 = -0.35$. Next we set $T = \exp\{1 + \mu + \exp(-0.5\mu^2)\epsilon\}$ and we take $\epsilon = \log(\epsilon^*)/\sqrt{\mathrm{var}(\epsilon^*)}$, where ϵ^* follows Weibull(k, λ), so that $\mathrm{pr}(\epsilon^* > 5) = 0.79$ and $\mathrm{pr}(\epsilon^* > 10) = 0.66$, where 0.79 (0.66) was the median 5-year (10-year) survival probability across all groups from the SEER data. Note that $\mathrm{pr}(\epsilon^* > r) = \exp\{(r/\lambda)^k\}$. Hence we get $k = 0.82$ and $\lambda = 29.27$, and $\sqrt{\mathrm{var}(\epsilon^*)} = 1.57$. We set the censoring variable $C = Z_1 + Z_2 + U_C$, where $U_C \sim$ Uniform$(0, K)$ distribution, and choose K such that on average 30% observations are right censored. This simulation design is henceforth referred as scenario 1.

Method of analyses: We simulated 500 such datasets by repeating the above procedure, and each data set was analyzed by the three model based methods: 1) the Cox model with the LASSO variable selection technique (Cox-LASSO) 2) the parametric AFT model with generalized F distribution for the residual, and 3) the proposed method referred to as AFT-Bayes-LASSO estimator. Also, the Kaplan-Meier method was applied to each group separately to estimate survival probabilities. The Cox-LASSO approach was implemented using R package `glmnet` [22], and there we obtained the LASSO-parameter λ by performing the 10-fold cross validation for variable selection. We implemented the parametric AFT model using R package `flexsurv` [23] which adopts a flexible regression approach for survival models. Following a referee's suggestion we considered a generalized F distribution for the parametric AFT model that allowed more flexibility in capturing different patterns of the survival time. Since `flexsurv` does not conduct any variable selection, we harness the existing functions for flexible regression with the classical variable selection procedures: forward selection and backward elimination. For each of the 2 stepwise procedures, we further considered 3 frequently used criteria: AIC, BIC, and p-value with the commonly adopted threshold of 0.05 for inclusion or exclusion.

For the proposed method, we took $N = 150$ which yields $\mathcal{L}_1$ distance in a Dirichlet process approximation error 2.29×10^{-9} ($\sim 4n\exp\{-(N-1)/\alpha\}$) for $n = 5,000$ and $\alpha = 5$ [24]. In the simulation and in the real data analysis, the variance was modeled as a quadratic function of the mean (i.e., $L = 2$), and set the hyperparameters $r = 0.01$ and $\delta = 0.01$ involved in (1), so that the prior mean and variance of λ^2 are 1 and 100, respectively. Also, we set $\tau_\eta^2 = 0.001$, $a_\alpha = b_\alpha = 1$, and initialize $\zeta = 1$. For the hyperparameters we set $\psi_1 = 0$, $\psi_2 = \psi_4 = 2.5$, $\psi_3 = \psi_5 = 1$ that result in a quite flexible base probability model under the finite-dimensional Dirichlet process. For each simulated data set, we ran 3 MCMC chains, each with $M = 20,000$ iterations. To compute one chain on a 2.66GHz Oct-core Intel Xeon E7-8837 processor it required approximately 44 minutes. The convergence was concluded after a burn-in period of $15,000$. We then sampled at every 10th iteration to form the posterior samples with total size $1,500$ (500 posterior samples from each of three chains) for posterior inference. In the end, we recorded the average of the posterior means, the standard deviation of the posterior means for each of the β-parameters, and the 95% credible interval for each parameter– based on which we declare if a variable is statistically significant or not.

Results: The performance of the aforementioned methods are shown in Figure 1, which visualizes the number of false positives (i.e., the number of significant βs that are actually null) and false negatives (i.e., the number of non-significant β that are actually non-null) using box-plots. The results indicate that the Cox-LASSO approach tends to select more variables and hence gives high proportion of false positives. A possible reason might be that in the simulation study we have generated data according to an AFT structure, instead of using a proportional hazard model structure.

Although a generalized F distribution has some flexibility in distribution assumptions, we found its implementation in R software can suffer a high chance of failure either in optimization of the likelihood functions, or in the computation of the Hessian matrix, for the p-value based variable selection procedures. We found the computational issue became more severe as the percentage of the censoring cases increased, even when we used some reasonable initialization of the parameters by fitting reduced models (with less number of parameters). The p-value based forward selection (GenF-f-pval) turns out to outperform other variants of the parametric AFT model with the stepwise selection procedures. In general, these procedures suffer from high rates of false negatives for backward elimination, and high rates of false positives for forward selection.

Overall, the proposed method outperforms others when jointly considering the false negatives and false positives. Figure 1 indicates that the proposed method detects variables with significant or trivial effects, with relatively smaller rates for both false positives (4.4%) and false negatives (0%). Additionally, in Figure 2 we show the proportion of times each β-parameter is significantly different from zero in the proposed method.

For the sake of comparison, among several classes, we considered four classes $(Z_1 = Z_2 = Z_3 = 1, Z_4 = 0)$, $(Z_1 = Z_3 = Z_4 = 1, Z_2 = 0)$, $(Z_1 = Z_2 = Z_3 = Z_4 = 1)$, and $(Z_1 = Z_2 = Z_3 = 0, Z_4 = 1)$. Under our choice of $(\beta_1, \beta_2, \beta_3, \beta_4) = (0.5, 0.5, 0.35, -0.35)$, the first class has the highest average survival rate, in particular near the starting period, while the last class, in contrast, has the smallest mean survival rate. The other two classes are intermediate groups. For these classes we estimate the survival probability and 95% confidence bands based on the 500 simulations. They are presented in Figure 3 along with the true survival probabilities. The Kaplan Meier estimator fitted for each group can capture the true survival curve quite well with a higher level of uncertainty. In contrast, the proposed method allows information sharing across the groups resulting in a lower level of uncertainty (i.e., much narrower confidence bands) in the estimator. Notably, both the Cox-LASSO and the parametric AFT with generalized F distribution significantly deviate from the true survival curve, in particular for the extreme groups, $(Z_1 = Z_2 = Z_3 = 1, Z_4 = 0)$ and $(Z_1 = Z_2 = Z_3 = 0, Z_4 = 1)$. The results suggest that the proposed method can recover well the underlying survival probabilities given a large proportion of censored cases, and has a better performance compared to the commonly adopted methods.

We also conducted another simulation study (henceforth referred as scenario 2) with the following design. We simulated Z_j's in the same way as the previous scenario, but took $\beta_1 = \beta_2 = 0.2$, $\beta_3 = -\beta_4 = 0.1$, and $T = \exp\{1 + \mu + \exp(0.5 + \mu^2)\epsilon\}$, $C = Z_1 + Z_2 + U_C$, $U_C \sim \text{Uniform}(0, K)$, where K is chosen so that the data contain a high proportion (50%) of right censored cases. This scenario has more censored subjects, and unlike the previous scenario, this study has very low signal-to-noise ratio (i.e., $\mu / \exp(0.5 + \mu^2)$ is low). We provide box-plots of false positives and false negatives in Figure 4. The proposed method again outperforms other competitors in selecting important variables with relatively smaller rates for both false positives and false negatives, which are 3.2% and 0.6%, respectively. Also, Figure 5 shows the proportion of times each β-parameter is significantly different from zero in the proposed method. Finally, the estimated survival probabilities for the four aforementioned classes are given in Figure 6. The proposed method consistently provides more accurate estimates with narrower confidence bands that cover the true survival curves compared to the competing methods. The Cox-LASSO and the parametric AFT model also provide accurate estimates under scenario 2 where the simulated cases have generally higher survival rates. However, they both deviate from the true curve for the risky group (bottom-right panel). Overall, the proposed method shows very robust performance towards a large proportion of censored cases and a low signal-to-noise situation in extracting effective predictors and estimating the survival probabilities. For both scenarios we estimate the bias and root mean squared error of the regression parameter estimators under the three model based approaches. They are presented in Tables 1 and 2 of the supplementary materials. The results in Table 1 indicate a negligible bias for all 20 estimators in the proposed method. Although the results in Table 2, where the signal-to-noise ratio is small, show somewhat larger bias for the non-zero coefficients in the proposed method, the performance of the proposed method is again much better than the other approaches.

6 Analysis of the SEER breast cancer data

Data overview: We analyzed the survival time of female breast cancer patients using the SEER public use data (SEER 1973-2003 Public-Use CD). Following somewhat similar criteria as [2], we considered only the subjects identified through autopsy and death certificate, while excluding: 1) male breast cancer subjects; 2) the subjects for whom the breast cancer was not primary; 3) the subjects with unknown race and race other than Black and White, 4) the subjects with unknown survival time; 5) the subjects with an age at diagnosis less than 20 years; 6) the subjects with no microscopic confirmation and sarcomas; 7) the subjects with unknown grade; 8) the subjects with unknown stage; 9) the subjects with unknown ER status; and 10) the subjects with unknown PR status. Moreover, we considered subjects who were diagnosed between January 1, 1990 and December 31, 2000, because for breast cancer patients, ER and PR status information is only available from January 1, 1990.

The predictor variables consisted of four prognostic factors: stage with 6 categories, I, IIA, IIB, IIIA, IIIB, IV (we excluded stage 0 as there was only one such case following the preprocessing rule);

grade with 4 categories 1, 2, 3, 4; ER status with 2 categories, *positive* and *negative*; PR status with 2 categories *positive* and *negative*; and two demographic factors: race with 2 categories, *Black* and *White*; and age at diagnosis with 7 categories, *20-29*, *30-39*, *40-49*, *50-59*, *60-69*, *70-84*, *85 years and onwards.* Thus, the main effects and two-factor interactions lead to $6 + \binom{6}{2} = 6 + 15 = 21$ factors. Since all 6 factors are categorical variables, the number of regression parameters (number of components of $\boldsymbol{\beta}$) involved in the mean function is much larger than 21, and it is

$$\begin{aligned} q &= \underbrace{5+3+1+1+1+6}_{\text{main effects}} + \underbrace{5(3+1+1+1+6)+3(1+1+1+6)+1(1+1+6)+1(1+6)+1(6)}_{\text{two factor interactions}} \\ &= 125. \end{aligned}$$

This large dimension of $\boldsymbol{Z}$ clearly indicates the necessity of a variable selection method.

After the aforementioned exclusion from the original 224, 444 female subjects who were diagnosed with breast cancer during 1990–2000 our analyzed data set included $n = 92,147$ subjects. Our model includes $q = 125$ regression parameters from $p = 6$ factors. Define the survival time T as the time to death from the time of diagnosis calculated in months. The latest time of diagnosis, i.e., the last month for samples entering the study, was December, 2000. However, we observed V which is the time of death, date last known to be alive, or follow-up cutoff date (12/31/2003) whichever occurred first from the date of diagnosis. We present the 5-year survival statistics: approximately 20% of subjects died within 5 years, and 59% of subjects survived at least 5 years; the remaining 21% of subjects were censored within 5 years from the time of diagnosis. We also report the 10-year survival statistics: about 28.2% of subjects died within 10 years, 14.8% of subjects survived 10 years; and the remaining 57% of subjects were censored within 10 years from the time of diagnosis.

The age group 70-84 years had the smallest proportion of censored cases. One of the goals of the analysis is to estimate the survival probabilities for each group defined by the prognostic and demographic factors, and also identify important predictors of survival time.

Method of analyses: For the proposed method (AFT-Bayes-LASSO), we used the same priors that were used in the simulation study. Additionally, to assess the necessity of the variance modeling, we re-analyzed the data by setting $\gamma_1 = \gamma_2 = 0$ while everything else was the same as the previous analysis. We label this special case as no variance modeling, AFT-Bayes-LASSO (noVar). Furthermore, we analyzed the data using the Cox-LASSO approach, and the parametric AFT model with generalized F distribution. For the parametric AFT model we adopted a p-value based forward-selection method (GenF-f-pval) that outperformed the other stepwise variable selection techniques in our simulation study. Additionally, we estimated survival probabilities using the Kaplan-Meier method for each group separately.

The R package `glmnet` for the Cox-LASSO approach does not produce a standard error of the estimator. Therefore, for the sake of comparison, we used a bootstrap resampling method (with 500 bootstrap samples) to compute the standard error of the parameters and 95% pointwise confidence intervals for the survival probabilities. Note that each bootstrap sample may result in different sets of selected variables, and not-selected variables are set to zero. In calculating standard errors we included both zero and non-zero estimates.

Results and discussions: Figure 7 shows the posterior mean and 95% credible interval of the parameters under the proposed method. Compared to the AFT-Bayes-LASSO (noVar) approach where $\gamma_1 = \gamma_2 = 0$, AFT-Bayes-LASSO yielded much narrower credible intervals for the parameters. We also obtained the following posterior mean (95% credible interval) for the $\boldsymbol{\gamma}$-parameters, $\widehat{\gamma}_1 = 0.08(0.06, 0.11)$, $\widehat{\gamma}_2 = 0.20(0.17, 0.21)$ for the AFT-Bayes-LASSO method. This result clearly indicates γ_1 and γ_2 are statistically significantly different from 0. Secondly, we calculated the Bayes factor to compare the two nested model fits. The Bayes factor with a hugely large value favored the model where the variance was modeled as a function of the mean. These facts support the necessity of variance modeling. For the Cox-LASSO and GenF-f-pval methods, we only show those selected β_j's with 95% confidence bands.

Figure 7 shows that, out of the 125 β_j's, 71 (56.8%) were significant under our proposed AFT-

Bayes-LASSO approach. This is similar to the output from GenF-f-pval, which selected 63 (50.4%) β_j's. On the other hand, Cox-LASSO selected as many as 110 (88%) β_j's, which may not be quite sensible. This also echoes with our simulation results that Cox-LASSO has a higher rate of false positives. Although the signs of the main effect estimates were somewhat consistent across the methods, occasionally, for some of the interaction terms, the signs of the three estimates varied across the methods. We found that GenF-f-pval generally agrees with AFT-Bayes-LASSO (noVar) due to the fact that they are both AFT models, one with a parametric model for the residual and the other with a nonparametric model for the residual. Many β_j's turned out to be statistically significant under Cox-LASSO, especially some interaction terms between the stage and age groups that are largely non-significant in the other approaches. For the sake of completeness, in Table 3 of the supplementary materials, we provide the estimate and credible/confidence interval for each of 125 regression parameters under the different methods. Finally, in light of the simulation results, the parameter estimates under our AFT-Bayes-LASSO deemed to be more trustworthy.

It took more than a week to analyze the data with the GenF-f-pval method. The computing time for the Cox-LASSO method strongly depends on the specified convergence threshold for coordinate descent and the penalty parameter. Although the default convergence threshold 10^{-7} did not work, a larger threshold of 10^{-4} worked and returned the model parameter estimates within an hour. Standard errors were calculated separately using a bootstrap method. For AFT-Bayes-LASSO, it took around 10 hours for each of the MCMC runs that proceed in parallel, to obtain the full results for the Bayesian inference.

Next, we compared the results under different approaches in estimating the survival probabilities for individual groups. Although the maximally observed survival time was 167 months (from January 1990 to December 2003) the whole study period, we hereby present survival curves up to 10 years (120 months) for each group (the National Cancer Institute only attempts to estimate 5-year survival probability).

There were $6 \times 4 \times 2 \times 2 \times 2 \times 7 = 1,344$ possible cross-classified groups (unique value of the vector Z_i) for the 6 factors. Out of the $1,344$ possible groups, 304 groups were empty without any subjects. However, our proposed method like any model based approach, allows for estimating the survival probabilities for all the 1,344 groups due to information sharing across all groups. Both the sample size and percentage of censored cases within 10 years can vary significantly across groups. We demonstrate our model's capability of estimating survival probabilities using several representative groups as shown in Figure 8, where we again compare our method with the other approaches.

The first row (call Type-A) includes the groups with a large number of cases and a high proportion of censored cases. These groups also have higher survival probabilities. The second row (referred to as Type-B) consists of the groups that have moderate sample sizes and percentages of censoring cases. The third row contains the groups with small sample size: the first two are risky groups with rather small survival rates, while the last two have higher censoring percentages and survival rates. We call the first two as "Type-C" groups which represent high risk groups, and call the last two together with the 4 groups in the last row, as "Type-D" groups, which have a small sample size but relatively higher survival rates than Type-C groups. Figure 8 indicates that for Type-A, B, D groups, the estimates based on the 3 model-based approaches generally agree and have narrower confidence bands than the Kaplan-Meier estimators, which have a high uncertainty under fewer observations. Nevertheless, for the risky Type-C group, the Cox-LASSO approach provides estimated curves that largely deviate from the other estimates. On the other hand, for Type-C, the survival estimates under the GenF-f-pval approach are somewhat smaller than the other estimates. Our AFT-Bayes-LASSO approach does not show any such pattern.

Finally, in Table 2, we present the estimated 5-year ($t_0 = 60$ months) survival probabilities along with the 95% credible intervals for several representative groups. Table 4 of the supplementary materials contains the estimated survival probabilities along with the 95% credible intervals for all $1,344$ groups.

7 Conclusions

We developed a flexible Bayesian method for variable selection in the AFT model. We applied it to analyzing SEER breast cancer data that enables the estimation of 5-year survival probabilities for different groups defined by levels of the prognostic factors. To the best of our knowledge, this is the first comprehensive analysis of SEER data which accounts for all recorded prognostic factors in the estimation of survival probabilities. The proposed method can also be applied to analyze a recent version of the SEER data where one may encounter a higher percentage of censoring. Although a higher percentage of censoring likely to result in a loss of efficiency, the relative performance of the methods expected to be the same based on our simulation results for the 50% censoring case.

We considered a linear regression model for the logarithm of the survival time. The residual term of the linear model is flexibly modeled as an approximate Dirichlet process mixture of normal distributions. Furthermore, the variance was modeled as a function of the mean. A variable selection technique has been employed in the mean function using LASSO. All these components make the approach very flexible. The use of the Bayesian method allowed us to estimate the parameters in this complex and flexible model. Simulation results indicated that the method worked quite well in estimating survival probabilities, and identifying important variables that are significantly associated with the time to event. The methods work when the signal-to-noise ratio is small and the percentage of censored cases is relatively high.

The proposed model allowed the joint estimation of the main effects with importance by borrowing information across all the groups according to different levels of the prognostic factors. Thereby for the real data, it allowed us to estimate the survival probabilities for groups even with few observations. Hence we were able to 1) estimate the survival probabilities for all the groups introduced by the prognostic factors and 2) provide the corresponding credible intervals as a measure of uncertainty.

The extensive and realistic simulation study indicates very good performance of the proposed approach even when the sample size was close to $100,000$. The computer code in MATLAB and R for our approach and the parametric AFT model with the stepwise selection procedure will be publicly available through our website.

Acknowledgment

The authors are grateful to two anonymous reviewers for their helpful comments and suggestions that led to a significant improvement of the manuscript. This research was supported by NIH grant R03CA176760.

References

1. Surveillance, Epidemiology, and End Results (SEER) Program (`http://seer.cancer.gov`); 2006. Public-Use Data (1973-2003). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2006, based on the November 2005 submission.

2. Ries LAG, Young JL, Keel GE, Eisner MP, Lin YD, Horner MJ. SEER Survival Monograph: Cancer Survival Among Adults: U.S. SEER Program, 1988-2001, Patient and Tumor Characteristic; 2007. National Cancer Institute, SEER Program, NIH Pub. No. 07-6215, Bethesda, MD, 2007.

3. Buckley J, James I. Linear regression with censored data. *Biometrika* 1979; **66**(3): 429–436.

4. Wei LJ, Ying ZL, Lin DY. Linear regression analysis of censored survival data based on rank tests. *Biometrika* 1990; **77**(4): 845–851.

5. Christensen R, Johnson W. Modeling accelerated failure time with a Dirichlet process. *Biometrika* 1988; **75**(4): 693–704.

6. Kuo L, Mallick BK. Bayesian semiparametric inference for the accelerated failure-time model. *The Canadian Journal of Statistics* 1997; **25**(4): 457–472.

7. Walker S, Mallick BK. A Bayesian Semiparametric accelerated failure time model. *Biometrics* 1999; **55**(2): 477–483.

8. Tibshirani R. The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, 1997; **16**: 385–395.

9. Knight K, Fu W. Asymptotics for Lasso-type estimators. *The Annals of Statistics* 2000; **28**: 1356–1378.

10. Fan J. Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* 2002; **96**: 1348–1360.

11. Zou H. The adaptive LASSO and its oracle properties. *J. Amer. Statist. Assoc.* 2006; **101**: 1418–1429.

12. Huang J, Sun T, Ying Z, Yu Y, Zhang C-H. Oracle inequalities for the LASSO in the Cox model. *Annals of Statistics*, 2013, **3**: 1142–1165.

13. Huang J, Ma S, Xie H. Regularized Estimation in the Accelerated Failure Time Model with High Dimensional Covariates. *Biometrics* 2006; **62**(3): 813–820.

14. Sha N, Tadesse MG, Vannucci M. Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* 2006; **22**(18): 2262–2268.

15. Lee KH. *Bayesian Variable Selection in Parametric and Semiparametric High Dimensional Survival Analysis*. PhD Thesis, University of Missouri, Colombia, 2011.

16. Chiou JM, Müller HG. Estimated estimating equations: semiparametric inference for clustered/longitudinal data. *Journal of the Royal Statistical Society, Series B* 2005; **67**(4): 531–553.

17. Kauermann G, Wegener M. Functional variance estimation using penalized splines with principal component analysis. *Statistics and Computing* 2011; **21**(2): 159–171.

18. Ghosh JK, Ramamoorthi RV. *Bayesian Nonparametrics*. New York: Springer, 2003.

19. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 1996; **58**(1): 267–288.

20. Park T, Casella G. The Bayesian Lasso. *Journal of the American Statistical Association* 2008; **103**(482): 681–686.

21. Roberts GO, Rosenthal JS. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 2009; **18**(2): 349–367.

22. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 2010; **33**(1): 1–22. http://www.jstatsoft.org/v33/i01/

23. Jackson C. flexsurv: Flexible Parametric Survival and Multi-State Models. 2015. R package version 0.6. http://CRAN.R-project.org/package=flexsurv

24. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* 2001; **96**(453): 161–173.

25. Gelfand AE, Hills SE, Racine-Poon A, Smith AFM. Illustration of Bayesian inference in Normal data models using Gibbs sampling. *Journal of the American Statistical Association* 1990; **85**(412): 972–985.

Appendix: Details of the MCMC steps

Step 1. Sample $\boldsymbol{\beta}$ from a multivariate Normal distribution with variance and mean

$$\begin{aligned}\Sigma^{\dagger} &= \left\{\tilde{\boldsymbol{Z}}^{T}\tilde{\boldsymbol{Z}} + \Lambda_0\right\}^{-1},\\ \mu^{\dagger} &= \Sigma^{\dagger}\boldsymbol{Z}^{T}\mu_0\end{aligned}$$

where $\tilde{\boldsymbol{Z}}$ is the $n \times q$ scaled design matrix with its ith row equal to the ith row of $\boldsymbol{Z}$ multiplied by $\{(v_i\theta_{i2})^{-1} + \tau_{\eta}^{-2}\}^{1/2}$, Λ_0 is a $q \times q$ diagonal matrix with the jth diagonal entry σ_j^{-2} for $1 \leq j \leq q$, μ_0 is an $n \times 1$ matrix with the ith entry $(v_i\theta_{i2})^{-1}\{\log(T_i^*) - \sqrt{v_i}\theta_{i1}\} + \tau_{\eta}^{-2}\eta_i$ for $1 \leq i \leq n$. In this case, $\boldsymbol{\beta}$ can be efficiently sampled by evaluating the Cholesky decomposition of the precision matrix $\tilde{\boldsymbol{Z}}^{T}\tilde{\boldsymbol{Z}} + \Lambda_0 = \mathcal{A}\mathcal{A}^T$, which follows that $\Sigma^{\dagger} = (\mathcal{A}^{-1})^T\mathcal{A}^{-1}$, and a desired sample has the form $(\mathcal{A}^{-1})^T Z_0 + \mu^{\dagger} = (\mathcal{A}^{-1})^T(Z_0 + \mathcal{A}^{-1}\boldsymbol{Z}^T\mu_0)$ where Z_0 is a q-variate standard normal random variable, hence the sampling procedure involves solving two linear triangular systems:

- solve $\mathcal{A}\boldsymbol{x}_0 = \boldsymbol{Z}^T\mu_0$ for $\boldsymbol{x}_0$, and
- solve $\mathcal{A}^T\boldsymbol{\beta} = (Z_0 + \boldsymbol{x}_0)$ for $\boldsymbol{\beta}$

which yields a desired sample $\boldsymbol{\beta}$.

Step 2. Define $u_j = 1/\sigma_j^2$, then sample u_j from the inverse-Gaussian distribution

$$\pi(u_j|\text{rest}) \quad \propto \quad \frac{1}{u_j^{3/2}}\exp\left\{-\lambda^2\frac{(u_j - |\lambda/\beta_j|)^2}{2u_j|\lambda/\beta_j|^2}\right\},\ j = 1, \cdots, q.$$

Step 3. Sample λ^2 from

$$\pi(\lambda^2|\text{rest}) \propto (\lambda^2)^{r+q-1}\exp\{-(\delta + \sum_{j=1}^{q}\frac{\sigma_j^2}{2})\lambda^2\}.$$

Step 4. Sample γ_l from the following conditional distribution

$$\pi(\gamma_l|\text{rest}) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{1}{\theta_{i2}}\left\{\frac{\log(T_i^*) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{\sqrt{\exp(\sum_{l=1}^{L}\gamma_l\eta_i^l)}} - \theta_{i1}\right\}^2 - \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{L}\gamma_l\eta_i^l - \frac{(\gamma_l - m_{\gamma l})^2}{2\sigma_{\gamma_l}^2}\right],$$

for $l = 1, \ldots, L$. For this step we shall adopt the Metropolis-Hastings algorithm.

Step 5. When $\Delta_i = 0$ following Gelfand et al. [25] we resample T_i^* as

$$\log(T_i^*) \;=\; \boldsymbol{Z}_i^T\boldsymbol{\beta} + \sqrt{v_i}\theta_{i1} + \sqrt{v_i\theta_{i2}}\Phi^{-1}\left\{(1-R)\Phi\left(\frac{\log(V_i) - \boldsymbol{Z}_i^T\boldsymbol{\beta} - \sqrt{v_i}\theta_{i1}}{\sqrt{v_i\theta_{i2}}}\right)+R\right\},$$

where $R \sim$ Uniform$(0,1)$, and Φ and Φ^{-1} are the CDF and the inverse of the CDF of the standard normal distribution, respectively.

Step 6. Sample η_i from the following conditional distribution

$$\pi(\eta_i|\text{rest}) \propto \exp\left[-\frac{1}{2\theta_{i2}}\left\{\frac{\log(T_i^*) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{\sqrt{\exp(\sum_{l=1}^{L}\gamma_l\eta_i^l)}} - \theta_{i1}\right\}^2 - \frac{1}{2}\sum_{l=1}^{L}\gamma_l\eta_i^l - \frac{(\eta_i - \boldsymbol{Z}_i^T\boldsymbol{\beta})^2}{2\tau_\eta^2}\right], \text{for } i = 1, \ldots, n.$$

Step 7. Sample the configuration indicators as follows. Sample s_i from $\pi(s_i|\text{rest}) \sim$ Multinomial$(p_{i1}^*, \ldots, p_{iN}^*)$, where $(p_{i1}^*, \cdots, p_{iN}^*) = K(p_1\phi[\{\log(T_i^*) - \boldsymbol{Z}_i^T\boldsymbol{\beta} - \sqrt{v_i}\vartheta_{11}\}, 0, v_i\vartheta_{12}], \cdots, p_N\phi[\{\log(T_i^*) - \boldsymbol{Z}_i^T\boldsymbol{\beta} - \sqrt{v_i}\vartheta_{N1}\}, 0, v_i\vartheta_{N2}])$, where K is a normalizing constant. If the new proposal of s_i is j, update $m_{s_i} = m_{s_i} - 1$, $m_j = m_j + 1$, $s_i = j$, $\theta_i = \vartheta_j$.

Step 8. Sample $(p_1, \cdots, p_N)$ from its conditional distribution, Dirichlet$(\alpha/N + m_1, \cdots, \alpha/N + m_N)$.

Step 9. Update $\vartheta_1, \ldots, \vartheta_N$ as follows. If $m_j > 0$, sample ϑ_j from

$$\pi(\vartheta_j|\text{rest}) \propto h(\vartheta_j|\psi)\prod_{i:s_i=j}\frac{1}{\sqrt{v_i}}\phi[\{\log(T_i^*) - \boldsymbol{Z}_i^T\boldsymbol{\beta} - \sqrt{v_i}\vartheta_{j1}), 0, v_i\vartheta_{j2}]$$

otherwise $\vartheta_j \sim H(\vartheta_j|\psi)$ for $j = 1, \cdots, N$. In particular, when $m_j > 0$, sample ϑ_{j1} from the Normal distribution with variance and mean

$$\begin{aligned}\sigma_\vartheta^2 &= \left(\frac{1}{\zeta\vartheta_{j2}} + \frac{m_j}{\vartheta_{j2}}\right)^{-1}, \\ \mu_\vartheta &= \sigma_\vartheta^2\left(\frac{\psi_1}{\zeta\vartheta_{j2}} + \sum_{i:s_i=j}\frac{\log(T_i^*) - \boldsymbol{Z}_i^T\boldsymbol{\beta}}{\sqrt{v_i}\vartheta_{j2}}\right),\end{aligned}$$

respectively, and sample ϑ_{j2} from the Inverse-Gamma distribution with shape and scale

$$a_\vartheta = \left(\psi_2 + \frac{1}{2} + \frac{m_j}{2}\right),$$

$$b_\vartheta = \left[\frac{1}{\psi_3} + \frac{(\vartheta_{j1} - \psi_1)^2}{2\zeta} + \sum_{i:s_i=j} \frac{\{\log(T_i^*) - \boldsymbol{Z}_i^T \boldsymbol{\beta} - \sqrt{v_i}\vartheta_{j1}\}^2}{2v_i}\right]^{-1},$$

respectively. Finally, sample ζ from the Inverse-Gamma distribution with shape and scale

$$\begin{aligned} a_\zeta &= \psi_4 + 0.5 \sum_{j=1}^N I(m_j > 0), \\ b_\zeta &= \left\{\sum_{j:m_j>0} \frac{(\vartheta_{j1} - \psi_1)^2}{2\vartheta_{j2}} + \frac{1}{\psi_5}\right\}^{-1}, \end{aligned}$$

respectively.

Step 10. Since the prior of α is Gamma(a_α, b_α), we sample α from

$$\pi(\alpha|\text{rest}) \propto \frac{\Gamma(\alpha)}{\{\Gamma(\alpha/N)\}^N} p_1^{\alpha/N-1} \cdots p_N^{\alpha/N-1} \alpha^{a_\alpha - 1} \exp(-\alpha/b_\alpha).$$

To draw α we use a Metropolis-Hastings algorithm with $\pi(\alpha)$, the prior density as the proposal density. Suppose that at the $(t+1)$th iteration we draw $\alpha^{(new)}$ from $\pi(\alpha)$, then

$$\alpha^{(t+1)} = \left\{ \begin{array}{l} \alpha^{(new)} \text{ with probability } \rho(\alpha^{(new)}, \alpha^{(t)}) \\ \alpha^{(t)} \text{ otherwise.} \end{array} \right. ,$$

where

$$\rho(\alpha^{(new)}, \alpha^{(t)}) = \min\left\{1, \frac{p_1^{\alpha^{(new)}/N-1} \times \cdots \times p_N^{\alpha^{(new)}/N-1} \Gamma(\alpha^{(new)}) / \{\Gamma(\alpha^{(new)}/N)\}^N}{p_1^{\alpha^{(t)}/N-1} \times \cdots \times p_N^{\alpha^{(t)}/N-1} \Gamma(\alpha^{(t)}) / \{\Gamma(\alpha^{(t)}/N)\}^N}\right\}.$$

Table 1: Estimates of the statistically significant β parameters along with the 95% credible intervals. Stage IV, grade 3, ER positive, PR positive, White race, and age group of diagnosis 70 to 84 years were used as the reference category of the respective variables.

Variable	Estimate	95% credible interval	
		Lower	Upper
Stage I	1.099	1.045	1.149
Stage IIA	0.935	0.880	0.990
Stage IIB	0.770	0.683	0.836
Stage IIIA	0.599	0.504	0.686
Stage IIIB	0.486	0.388	0.587
Grade 1	0.125	0.022	0.260
Grade 2	−0.067	−0.116	−0.011
Grade 4	−0.276	−0.385	−0.173
ER–	−0.737	−0.831	−0.663
PR–	−0.325	−0.389	−0.247
Black race	−0.416	−0.494	−0.316
Age of diag. 20–29	0.281	0.018	0.465
Age of diag. 30–39	0.560	0.376	0.723
Age of diag. 40–49	0.505	0.417	0.591
Age of diag. 50–59	0.359	0.277	0.422
Age of diag. 60–69	0.063	0.011	0.127
Age of diag. 85+	−0.845	−0.930	−0.688
Stage IIB × Grade 1	0.208	0.065	0.302
Stage IIIA × Grade 1	0.225	0.057	0.356
Stage I × Grade 2	0.146	0.074	0.200
Stage IIA × Grade 2	0.149	0.088	0.205
Stage IIB × Grade 2	0.197	0.125	0.260
Stage IIIA × Grade 2	0.218	0.120	0.276
Stage IIIB × Grade 2	0.211	0.132	0.304
Stage I × Grade 4	0.341	0.210	0.469
Stage IIA × Grade 4	0.262	0.145	0.411
Stage IIB × Grade 4	0.279	0.115	0.463
Stage IIIA × Grade 4	0.299	0.168	0.502
Stage IIIB × Grade 4	0.254	0.112	0.397
Stage I × ER–	0.658	0.585	0.777
Stage IIA × ER–	0.612	0.527	0.730
Stage IIB × ER–	0.566	0.494	0.660
Stage IIIA × ER–	0.519	0.374	0.645

Continued on the next page

Table 1 – continued from the previous page

Variable	Estimate	95% credible interval	
		Lower	Upper
Stage IIIB × ER–	0.474	0.353	0.590
Stage I × PR–	0.283	0.202	0.356
Stage IIA × PR–	0.229	0.140	0.321
Stage IIB × PR–	0.194	0.102	0.280
Stage IIIA × PR–	0.157	0.052	0.269
Stage IIIB × PR–	0.157	0.050	0.292
Stage I × Black race	0.345	0.244	0.411
Stage IIA × Black race	0.324	0.191	0.401
Stage IIB × Black race	0.318	0.229	0.405
Stage IIIA × Black race	0.255	0.110	0.380
Stage IIIB × Black race	0.181	0.111	0.300
Stage IIIB × Age of diag. 40–49	−0.212	−0.311	−0.054
Stage I × Age of diag. 50–59	0.133	0.047	0.201
Stage I × Age of diag. 60–69	0.304	0.218	0.355
Stage IIA × Age of diag. 60–69	0.283	0.170	0.359
Stage IIB × Age of diag. 60–69	0.204	0.105	0.254
Stage IIIA × Age of diag. 60–69	0.263	0.170	0.378
Stage IIIB × Age of diag. 60–69	0.131	0.049	0.216
Stage I × Age of diag. 85+	0.278	0.144	0.421
Stage IIA × Age of diag. 85+	0.263	0.112	0.384
Stage IIB × Age of diag. 85+	0.280	0.077	0.405
Stage IIIA × Age of diag. 85+	0.350	0.135	0.520
Stage IIIB × Age of diag. 85+	0.304	0.112	0.459
Grade 2 × PR–	−0.024	−0.052	−0.004
Grade 1 × Age of diag. 20–29	0.438	0.180	0.700
Grade 1 × Age of diag. 30–39	0.144	0.032	0.252
Grade 4 × Age of diag. 30–39	−0.071	−0.145	−0.003
Grade 1 × Age of diag. 40–49	0.169	0.124	0.230
Grade 2 × Age of diag. 40–49	0.065	0.039	0.111
Grade 1 × Age of diag. 50–59	0.061	0.029	0.097
Grade 2 × Age of diag. 50–59	0.064	0.036	0.091
ER– × PR–	0.041	0.014	0.074
ER– × Age of diag. 40–49	−0.049	−0.106	−0.009
ER– × Age of diag. 60–69	−0.026	−0.049	0.000
PR– × Age of diag. 20–29	0.164	0.044	0.296
PR– × Age of diag. 60–69	0.026	0.002	0.068
Black race × Age of diag. 40–49	−0.054	−0.086	−0.009
Black race × Age of diag. 60–69	−0.102	−0.173	−0.019

Table 2: Estimated 5 year survival probabilities along with the 95% credible intervals for some combination of prognostic factors and for all age groups for the diagnosis.

Stage	Grade	ER	PR	Race	Age of diagnosis						
					20-29	30-39	40-49	50-59	60-69	70-84	85+
I	1	+	+	W	0.956	0.960	0.964	0.950	0.919	0.819	0.553
					(0.933	(0.947	(0.961	(0.947	(0.914	(0.811	(0.529
					0.979)	0.969)	0.968)	0.953)	0.924)	0.828)	0.575)
I	1	+	+	B	0.947	0.945	0.949	0.930	0.874	0.784	0.500
					(0.910	(0.930	(0.940	(0.923	(0.852	(0.760	(0.438
					0.977)	0.959)	0.957)	0.940)	0.893)	0.809)	0.552)
I	2	+	+	W	0.831	0.937	0.948	0.944	0.917	0.800	0.497
					(0.778	(0.932	(0.944	(0.940	(0.912	(0.793	(0.475
					0.885)	0.942)	0.951)	0.947)	0.922)	0.808)	0.514)
I	2	+	+	B	0.808	0.914	0.928	0.923	0.873	0.767	0.450
					(0.733	(0.904	(0.919	(0.912	(0.855	(0.749	(0.381
					0.882)	0.925)	0.935)	0.935)	0.890)	0.790)	0.523)
I	3	+	+	W	0.839	0.923	0.926	0.920	0.893	0.767	0.469
					(0.794	(0.913	(0.921	(0.914	(0.887	(0.754	(0.431
					0.889)	0.932)	0.930)	0.926)	0.900)	0.779)	0.518)
I	3	+	+	B	0.821	0.898	0.900	0.895	0.842	0.735	0.428
					(0.750	(0.880	(0.888	(0.881	(0.822	(0.708	(0.353
					0.880)	0.911)	0.911)	0.907)	0.865)	0.757)	0.517)
IIA	2	+	+	W	0.822	0.895	0.929	0.908	0.873	0.730	0.401
					(0.724	(0.886	(0.924	(0.903	(0.867	(0.713	(0.383
					0.889)	0.903)	0.932)	0.913)	0.881)	0.743)	0.417)
IIA	2	+	+	B	0.789	0.853	0.897	0.871	0.804	0.680	0.349
					(0.663	(0.834	(0.885	(0.856	(0.771	(0.657	(0.296
					0.880)	0.871)	0.909)	0.891)	0.834)	0.707)	0.419)
IIB	2	+	+	W	0.740	0.840	0.879	0.861	0.807	0.671	0.352
					(0.647	(0.826	(0.871	(0.847	(0.797	(0.655	(0.320
					0.840)	0.855)	0.887)	0.870)	0.818)	0.689)	0.402)
IIB	2	+	+	B	0.698	0.781	0.828	0.808	0.715	0.614	0.302
					(0.561	(0.756	(0.795	(0.786	(0.669	(0.557	(0.238
					0.790)	0.810)	0.846)	0.834)	0.757)	0.654)	0.382)
IIB	3	+	+	W	0.727	0.791	0.815	0.791	0.739	0.601	0.305
					(0.635	(0.768	(0.800	(0.771	(0.721	(0.573	(0.280
					0.817)	0.811)	0.827)	0.816)	0.757)	0.626)	0.329)
IIB	3	+	+	B	0.689	0.726	0.752	0.728	0.637	0.546	0.263
					(0.592	(0.698	(0.703	(0.684	(0.596	(0.479	(0.207
					0.781)	0.751)	0.773)	0.751)	0.684)	0.578)	0.312)
IV	2	+	+	W	0.273	0.448	0.453	0.377	0.234	0.204	0.076
					(0.167	(0.371	(0.412	(0.334	(0.208	(0.183	(0.063
					0.399)	0.510)	0.491)	0.426)	0.255)	0.222)	0.089)
IV	2	+	+	B	0.151	0.235	0.235	0.192	0.112	0.113	0.075
					(0.097	(0.182	(0.194	(0.159	(0.095	(0.097	(0.059
					0.209)	0.300)	0.278)	0.224)	0.126)	0.130)	0.091)
IV	3	+	+	W	0.342	0.484	0.454	0.378	0.248	0.225	0.081
					(0.219	(0.406	(0.424	(0.350	(0.223	(0.208	(0.066
					0.440)	0.558)	0.495)	0.414)	0.269)	0.244)	0.096)

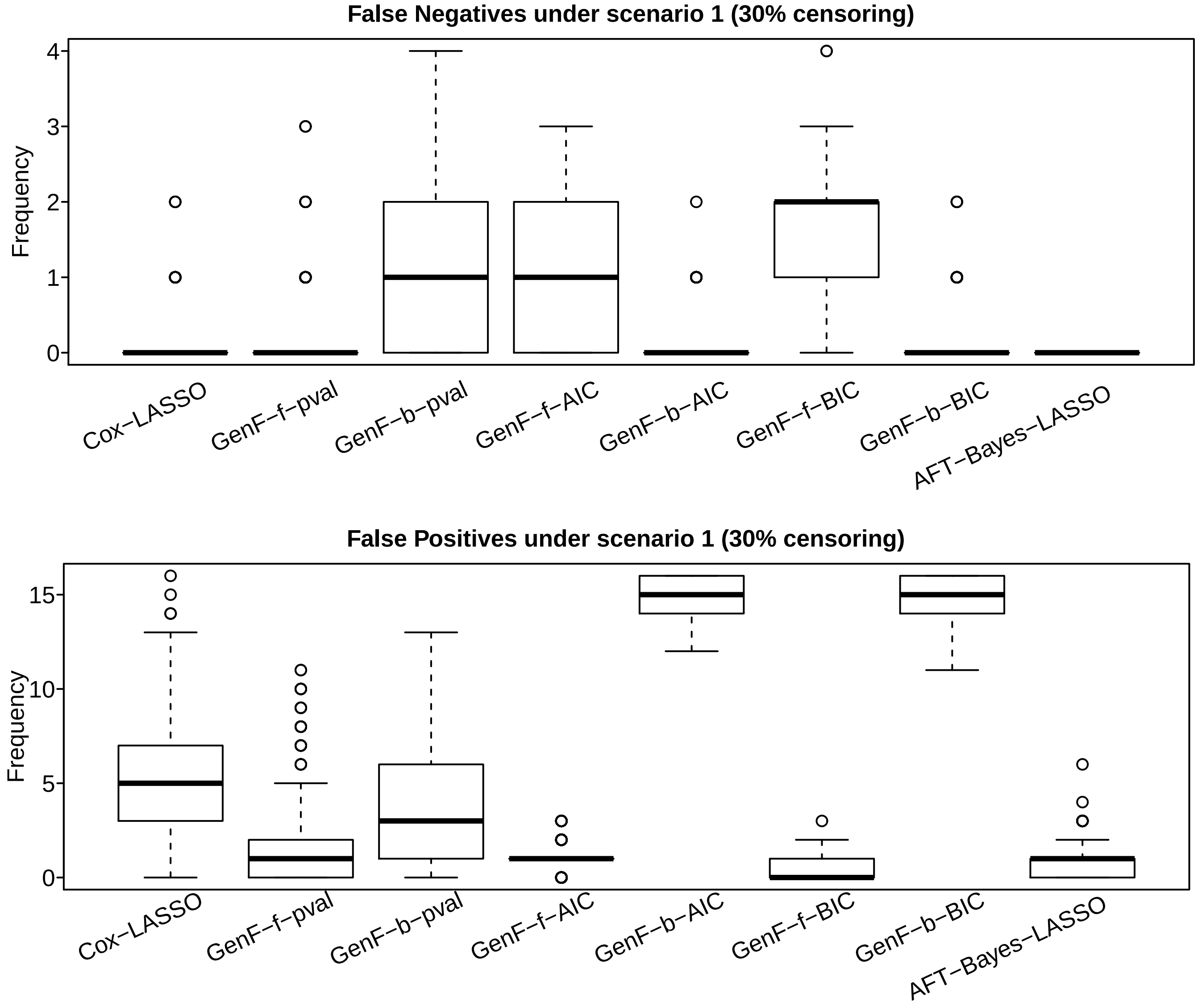

Figure 1: Box plot of false positives and false negatives based on 500 simulations under scenario 1. Here `GenF-f` and `GenF-b` stand for the parametric AFT model with generalized F distribution with the forward selection and backward elimination procedure, respectively, and `pval`, `AIC`, and `BIC` refer to the criteria measure used for the variable selection.

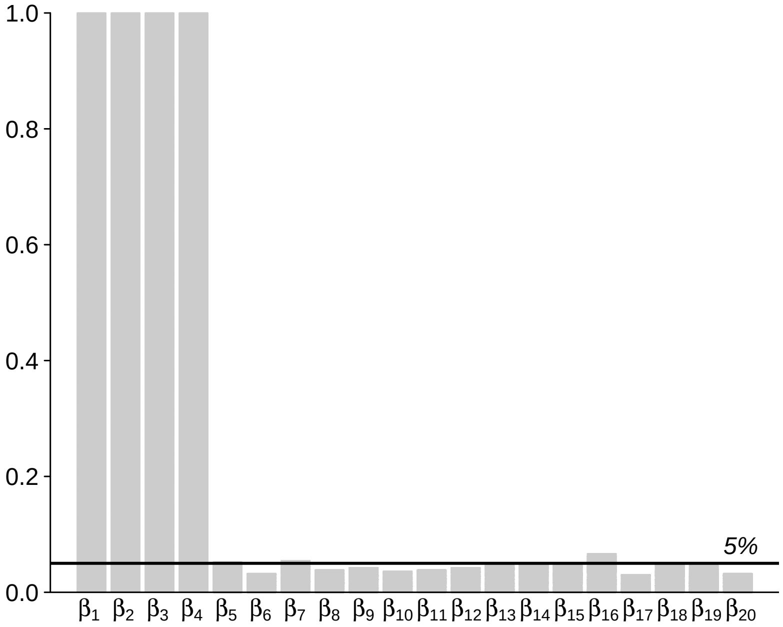

Figure 2: Bar diagram of the proportion of times each β parameter came out to significant based on 500 simulations under scenario 1.

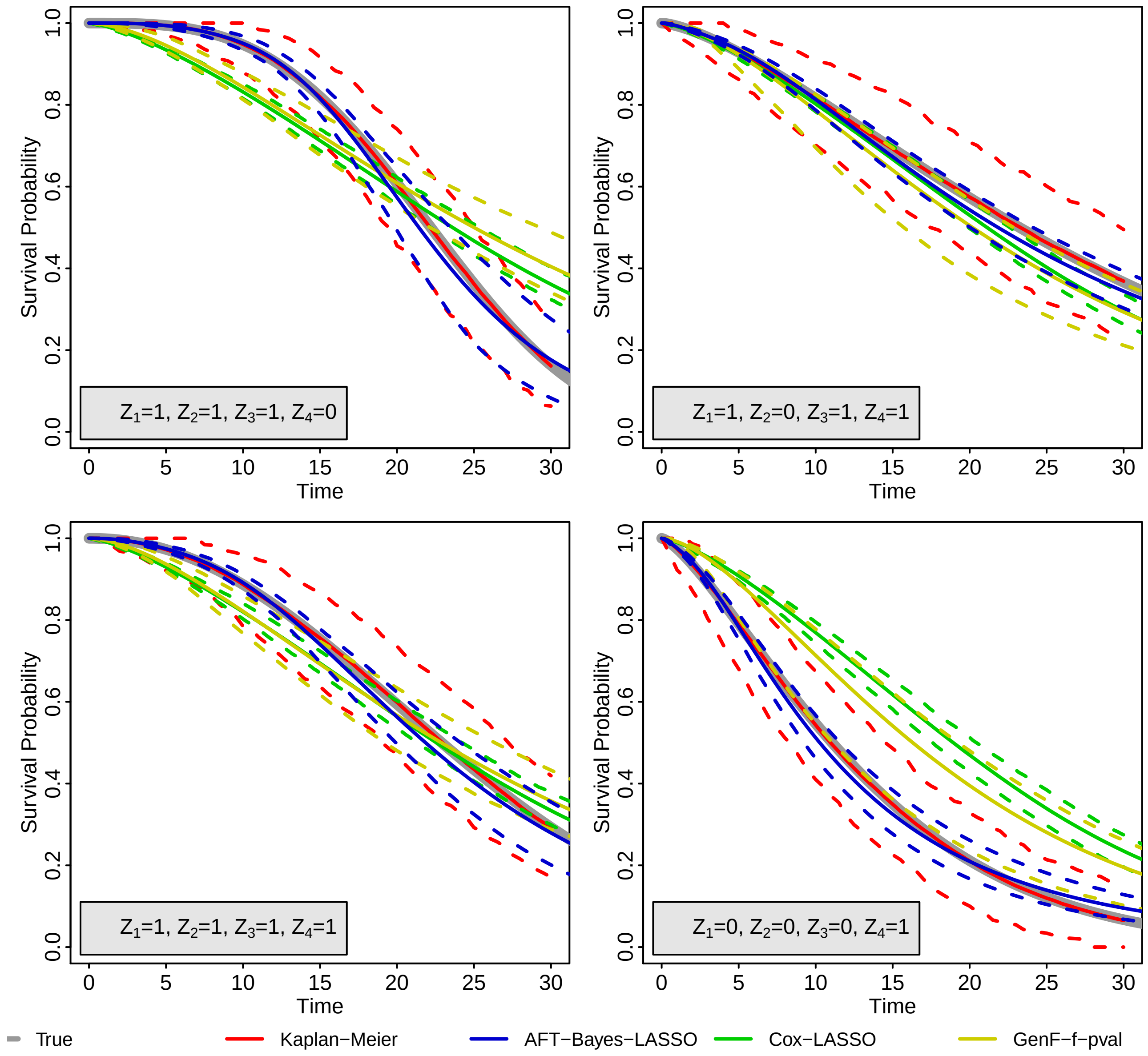

Figure 3: The average of the estimated survival probability, 95% pointwise credible interval based on the simulation study and the true survival probabilities under scenario 1 with 30% censoring cases.

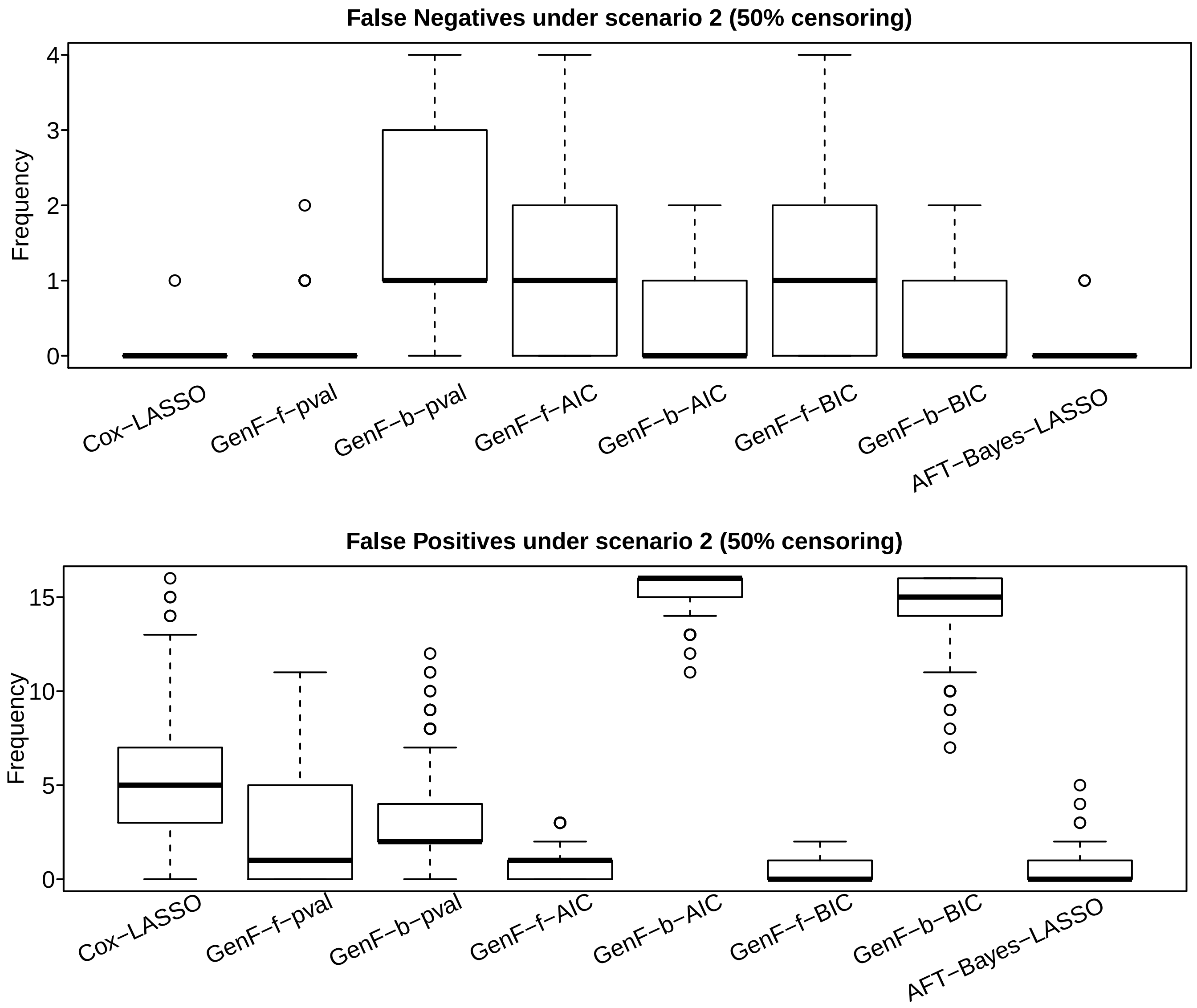

Figure 4: Box plot of false positives and false negatives based on 500 simulations under scenario 2. Here `GenF-f` and `GenF-b` stand for the parametric AFT model with generalized F distribution with the forward selection and backward elimination procedure, respectively, and `pval`, `AIC`, and `BIC` refer to the criteria measure used for the variable selection.

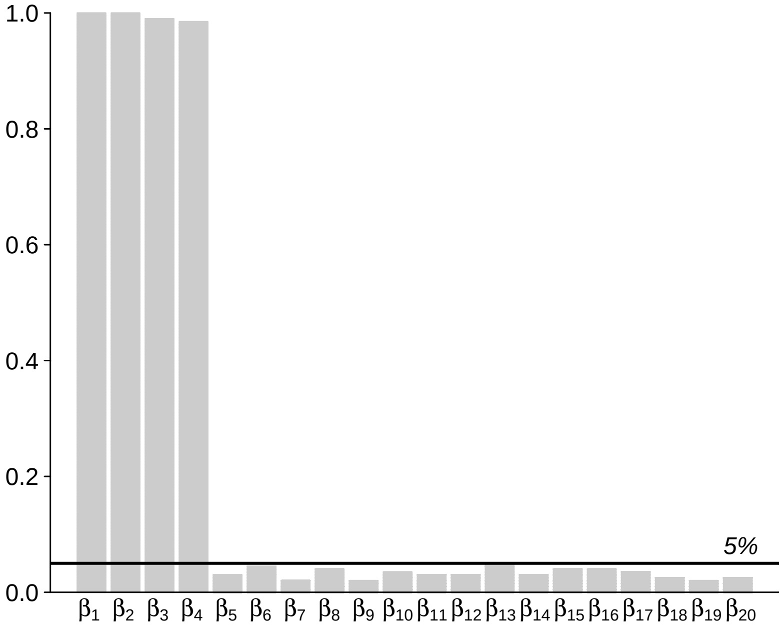

Figure 5: Bar diagram of the proportion of times each β parameter came out to significant based on 500 simulations under scenario 2.

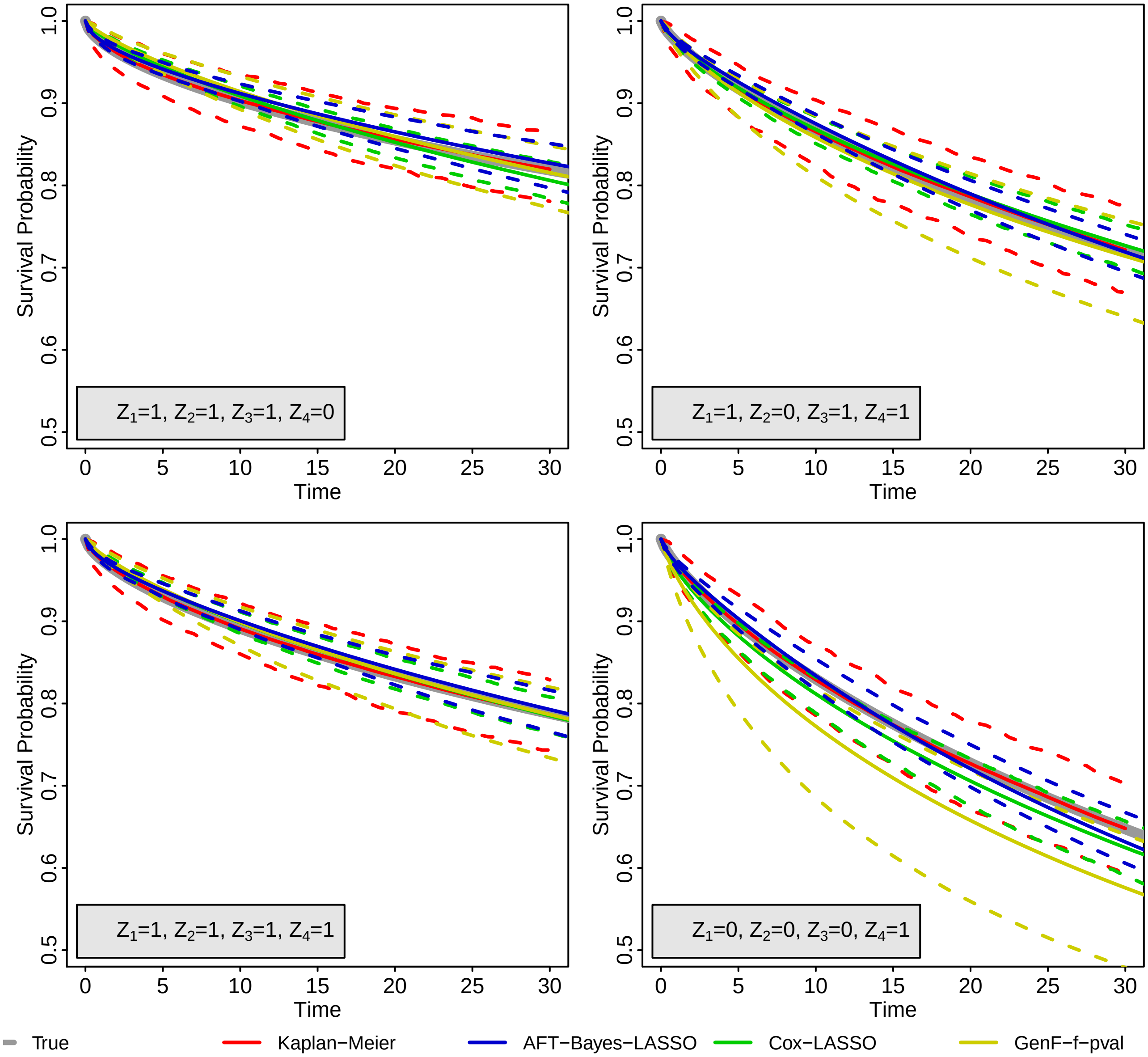

Figure 6: The average of the estimated survival probability, 95% pointwise credible interval based on the simulation study and the true survival probabilities under scenario 2 with 50% censoring cases.

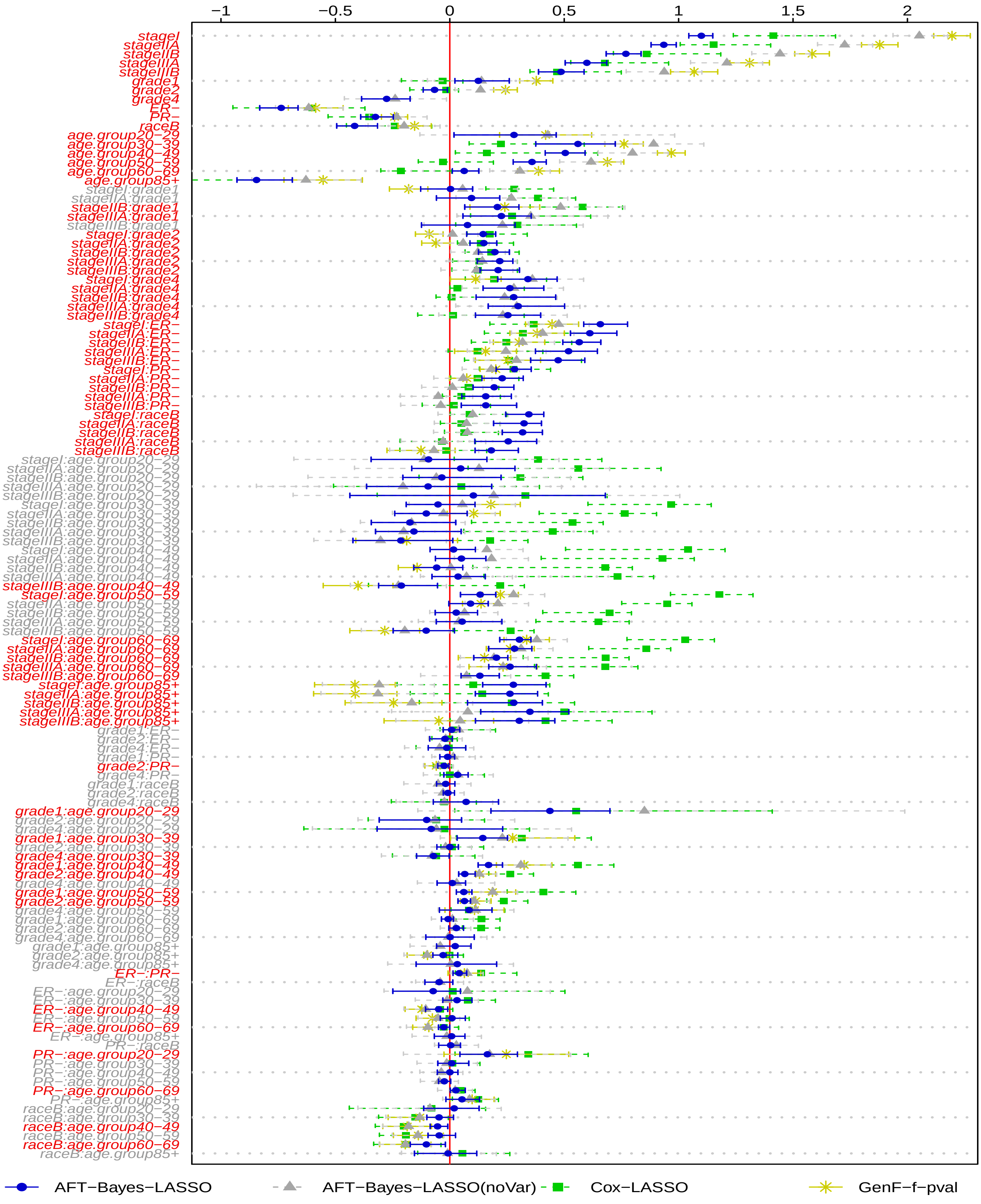

Figure 7: Plot of the estimate and 95% confidence/credible interval for $q = 125$ β_j parameters for the real data.

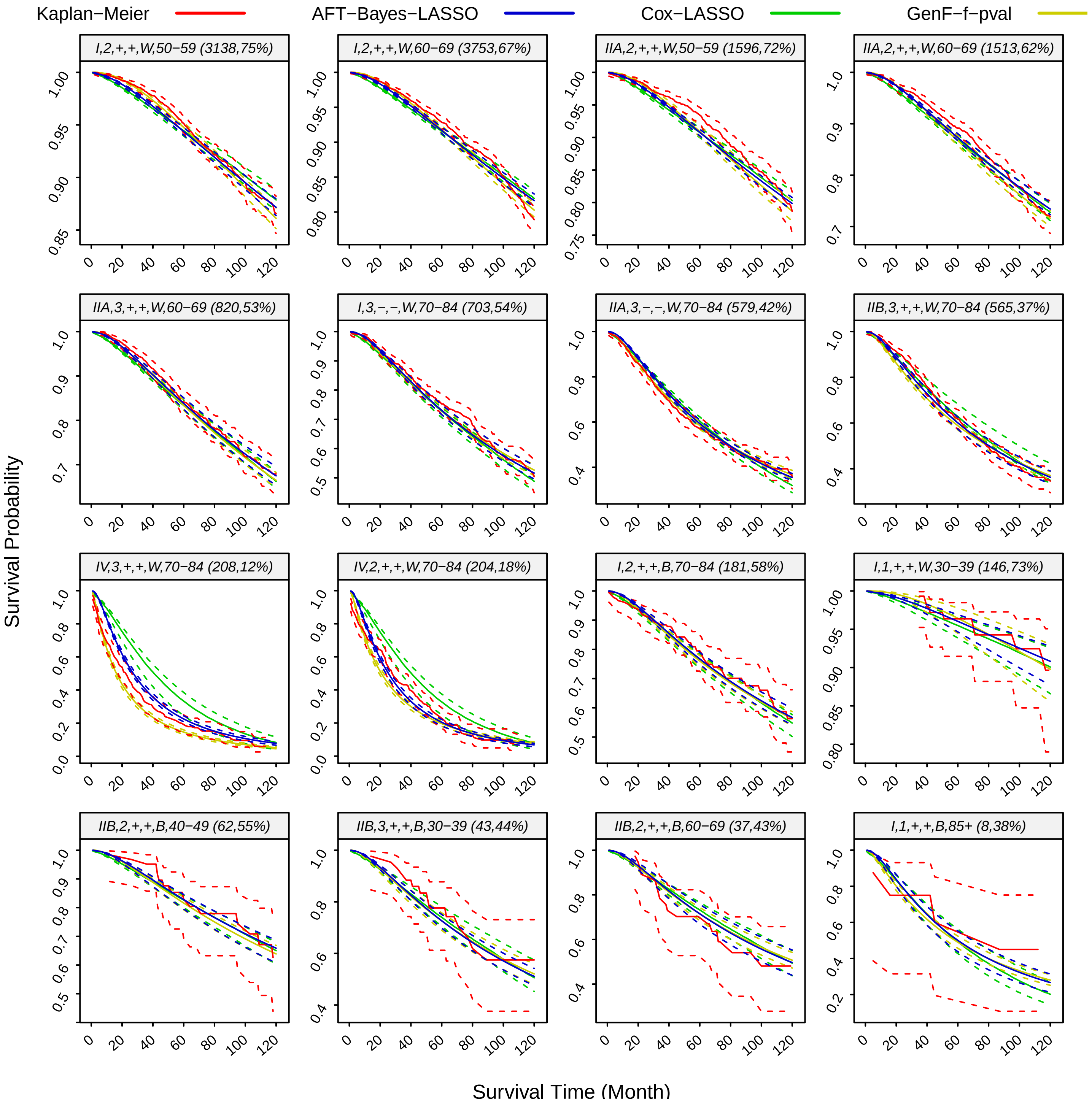

Figure 8: Estimated 10-year survival probabilities along with the 95% confidence/credible interval. The header of each group (panel) contains Stage, Grade, ER status, PR status, Race, Age group (Total number of cases, Percentage of censored cases within 10 years).