Received: 3 October 2018 | Accepted: 19 July 2019

DOI: 10.1111/biom.13132

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

BIOMETRIC PRACTICE

Bayesian data integration and variable selection for pan-cancer survival prediction using protein expression data

Arnab Kumar Maity[1] | Anirban Bhattacharya[2] | Bani K. Mallick[2] | Veerabhadran Baladandayuthapani[3]

[1]Early Clinical Development Oncology Statistics, Pfizer Inc., San Diego, California

[2]Department of Statistics, Texas A&M University, College Station, Texas

[3]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan

Correspondence
Arnab Kumar Maity, Early Clinical Development Oncology Statistics, Pfizer Inc., San Diego 92121, CA.
Email: arnab.maity@pfizer.com

Funding information
NIH, Grant/Award Numbers: R01CA194391, R21CA220299, R01CA160736; NSF, Grant/Award Numbers: DMS 1653404, DMS1463233

Abstract
Accurate prognostic prediction using molecular information is a challenging area of research, which is essential to develop precision medicine. In this paper, we develop translational models to identify major actionable proteins that are associated with clinical outcomes, like the survival time of patients. There are considerable statistical and computational challenges due to the large dimension of the problems. Furthermore, data are available for different tumor types; hence data integration for various tumors is desirable. Having censored survival outcomes escalates one more level of complexity in the inferential procedure. We develop Bayesian hierarchical survival models, which accommodate all the challenges mentioned here. We use the hierarchical Bayesian accelerated failure time model for survival regression. Furthermore, we assume sparse horseshoe prior distribution for the regression coefficients to identify the major proteomic drivers. We borrow strength across tumor groups by introducing a correlation structure among the prior distributions. The proposed methods have been used to analyze data from the recently curated “The Cancer Proteome Atlas” (TCPA), which contains reverse-phase protein arrays–based high-quality protein expression data as well as detailed clinical annotation, including survival times. Our simulation and the TCPA data analysis illustrate the efficacy of the proposed integrative model, which links different tumors with the correlated prior structures.

KEYWORDS
AFT regression, borrowing strength, horseshoe, pan-cancer model, TCPA

1 | INTRODUCTION

Accurate prognostic prediction using molecular profiles is an essential ingredient to develop precision medicine. Molecular profiling data emerging from high-throughput technologies can be used to search for new biomarkers and to develop accurate prognostic tools and predictive models. In this paper, we used proteomics data to develop similar translational models for identifying major actionable proteins that are associated with clinical outcomes, like the survival time of patients.

Direct analysis of high-dimensional proteomics data has received widespread attention because it represents a powerful approach to understand the pathophysiology and therapy of cancer, which cannot be achieved by analyses solely driven by genomics or transcriptomics (Li *et al.*, 2013; Akbani *et al.*, 2014). Baladandayuthapani *et al.* (2014) noted that many proteins are regulated by posttranslational modifications, such as phosphorylation or cleavage events, that are not detected by the analysis of DNA or RNA. Moreover, several studies have also demonstrated marked discordance between messenger RNA (mRNA) and protein expression levels, particularly for genes in kinase signaling and cell cycle regulation pathways (Shankavaram *et al.*, 2007). It has been demonstrated recently, in both cancer cell lines and tumors, that different genetic mutations in the same signaling pathway can result in significant differences in the quantitative activation levels of downstream pathway effectors (Park *et al.*, 2010). These observations support the suggestion that direct measurements are essential to measure protein activation. To fulfill this thrust, a new protein expression data visualization tool has been generated for 31 types of tumors using reverse-phase protein arrays (RPPAs) and have been uploaded in a user-friendly data portal, The Cancer Proteome Atlas (TCPA; Li *et al.*, 2017).

The primary source of the data motivating our methodological work comes from this TCPA. The current data covers more than 7500 tumor samples and signaling pathways in cancer such as P13K, MAPK, and mTOR. The 31 cancer types include bladder, breast, colon, brain, head and neck, kidney, lung (adenocarcinoma and squamous cell carcinoma), rectum, ovarian, and uterine cancers. Moreover, TCPA presents the tumor cell lines: 439 samples in four cell line sets, including both baseline and drug-treated cell lines. Figure 1 in Li *et al.* (2013) shows a detailed schematic of the TCPA data portal. The mRNA expression and DNA (copy number) were matched over large cohorts of well-characterized TCGA patient tumors. In addition, there is detailed clinical annotation that includes survival times and clinical subtype/stage information on the tumor samples. To the best of our knowledge, this represents the largest collection of cancer functional proteomics data with parallel genomic, transcriptomic, and clinical data currently available. For a detailed description of TCPA data portal, we refer the readers to Li *et al.* (2013) and the references therein.

The information available in TCPA are collected and presented for each tumor type, which is likely to be different and independent across tumor types. A schematic diagram of TCPA data is provided in Table 1. To accomplish the scientific goal of enhancing the statistical power of the inference procedure, integration of such data has been shown extremely useful; we cite Hamid *et al.* (2009) for a comprehensive review. For instance, Daemen *et al.* (2009) used kernel-based approaches to integrate the genomics data.

The availability of detailed matched proteomics data on hundreds of tumors collated by TCPA provides a major opportunity to develop an integrated picture of commonalities, differences, and emergent themes across tumor lineages. With this aim, we propose to develop methods to integrate pan-cancer data across tumor lineages. The overreaching goal of this pan-cancer effort is to provide a bird's-eye view of the functional proteome encompassing multiple tumor lineages, which may help to suggest potential targets that are applicable to disease subsets or across diseases. The pan-cancer analysis was launched by TCGA, which involves integration across tumor types and organs of origin to gain better analytical breadth (Weinstein *et al.*, 2013). Our interest in this paper is to develop a pan-cancer model, which will particularly link different tumor groups.

Each tumor group in the TCPA data portal consists of at least 189 proteins. Therefore, we must deal with high-dimensional statistical analysis after merging the data. Because of the easy interpretation, one seeks for a parsimonious model selection strategy, which can be achieved via several regularization techniques existing in the literature. For instance, there exist different penalized regressions, lasso (Tibshirani, 1996) and extensions of lasso, including adaptive lasso (Zou, 2006), and many others. The Bayesian variable selection literature is also rich; stochastic search variable selection (George and McCulloch, 1993), variable selection with shrinking and diffusing priors (Narisetty and He, 2014) to name a few. The prior specification is critical in any Bayesian analysis, if not essential. Hence, to achieve shrinkage and sparsity via prior formulation is generally of interest. Toward this end, Carvalho *et al.* (2010) proposed shrinkage estimation via the Horseshoe prior. Other propositions include Dirichlet-Laplace prior to Bhattacharya *et al.* (2015). However, the applications of these shrinkage priors in the setting of censored data are very limited until now. To our knowledge, Peltola *et al.* (2014) provided a comparison study using different shrinkage priors and concluded in favor of horseshoe prior in terms of the predictive

TABLE 1 The Cancer Proteome Atlas data structure

Tumor groups	Survival time of subjects	Measurements of protein expressions
Tumor 1	t_{11}	$x_{111}, x_{121}, \ldots, x_{1p}$
	t_{21}	$x_{211}, x_{221}, \ldots, x_{2p}$
	…	…
	$t_{n_1 1}$	$x_{n_1 11}, x_{n_1 21}, \ldots, x_{n_1 p}$
Tumor 2	t_{12}	$x_{112}, x_{122}, \ldots, x_{1p}$
	t_{22}	$x_{212}, x_{222}, \ldots, x_{2p}$
	…	…
	$t_{n_2 2}$	$x_{n_2 12}, x_{n_2 22}, \ldots, x_{n_2 p}$
…	…	…
Tumor k	t_{1k}	$x_{11k}, x_{12k}, \ldots, x_{1p}$
	t_{2k}	$x_{21k}, x_{22k}, \ldots, x_{2p}$
	…	…
	$t_{n_k k}$	$x_{n_k 1k}, x_{n_k 2k}, \ldots, x_{n_k p}$

abilities of the models. They considered a parametric Weibull model on the survival time.

In the settings of survival regressions, the straightforward implementation of these variable selection techniques may not be tenable, particularly due to the censored observations—which require additional sampling methods from the censored space. For example, Lee and Mallick (2004) assumed a linear model on the scale parameter of the Weibull survival regression or assumed a Gamma process on the Cox proportional hazard model (Cox, 1972), while Zhang *et al.* (2018) proposed a Dirichlet process prior on the accelerated failure time (AFT) model (Miller, 1976). In addition, following the data augmentation approach of Tanner and Wong (1984), the censored observations can be imputed (Bonato *et al.*, 2011). In contrast, the survival versions of penalized regularizations have also been discussed; important references include lasso (Tibshirani, 1997) and ridge regression (Li and Luan, 2002) in the context of Cox model. When developing the AFT model is of particular interest, then there exist several proposals in the literature; for example, Huang *et al.* (2006); Cai *et al.* (2009) derived some regularized versions including lasso and Huang and Ma (2010) discussed bridge regression, and (Wang and Song, 2011) developed adaptive lasso for AFT models. Furthermore, Khan and Shaw (2016, 2017) developed a class of adaptive elastic net techniques and synthesized techniques for variable selection in AFT models. However, they did not consider Bayesian settings in their research.

In this paper, we develop a Bayesian hierarchical AFT model. Unlike Cox proportional hazard model in which the covariates act multiplicatively on the hazard function, the AFT model considers the additive effect of the covariates on the log of survival responses, which results in an intuitive linear regression interpretation (Wei, 1992). In addition, when dealing with high-dimensional proteomics data, it has been reported to have the poor mixing in the Markov chain Monte Carlo (MCMC) chain in fitting a Cox model (Sha *et al.*, 2006). In Cox proportional hazard models, the regression parameters cannot be integrated out due to unavailability of any conjugate priors, thus requiring complex MCMC procedures. On the contrary, in our Bayesian log-normal AFT model, as we show in this article, most of the conditional distributions are available, and therefore, straightforward Gibbs sampling can be employed to update the chain.

Although the variable selection techniques described above have been useful to identify important features in genomics data, they have not been developed to incorporate the data integration procedures. On the contrary, the existing integration methods are not well-examined for high-dimensional pan-cancer settings for a single platform. In particular, in the presence of censored data, we observe a lack of a unified method which can be employed to deal with all these issues. Toward this end, we propose a Bayesian hierarchical model which fits a log-normal AFT model in each group. In addition, to achieve integration across human tumor groups, we model the mean parameter of the prior distributions on the coefficients to borrow strength across groups. To accomplish sparsity via the shrinkage priors, we place Horseshoe priors (or variants) on the coefficients. The full methodology has been combined and implemented in the R package *hsaft*. The resulting estimates are shown to be efficient compared to the existing regression methods run for each tumor group separately. In the end, selection of the most important proteins for the TCPA data is done using predictive survival curve. In particular, we use the Brier score (Brier, 1950) to asses the predictive performance of our fitted model.

A conceptually related field is multitasked survival analysis, which was introduced very recently in Li *et al.* (2016), Wang *et al.* (2017), and Liu *et al.* (2018). Among these, Wang *et al.* (2017) proposed a unified framework to integrate the multiple survival models for multiple tumor groups and provide a set of output models for each group. On the contrary, we propose a single model that accommodates all the tumor groups simultaneously. Furthermore, they worked with the Cox proportional hazard model while we employ the Bayesian analysis on the parametric AFT model.

The remainder of this article is organized as follows. In Section 2, we discuss the Bayesian log-normal AFT model, the prior formulation for borrowing strength, the sampling strategies from the posterior distribution space of the parameters, and the consistency property of the parameter estimator. In Section 3 we describe the survival prediction and the variable selection strategy from the MCMC samples. To show that the proposed method is superior to the individual analysis, we present simulation examples in Section 4. Finally, we illustrate the application of our proposed technique in TCPA data in Section 5 followed by a brief discussion in Section 6 to conclude this article.

2 | MODEL

From Table 1, it can be noted that the RPPA-based protein expression data in TCPA are expressed for several tumors; for instance, there are three types of kidney tumors, two types of lung cancers, and so on. The objective is to identify the major proteins common across the tumor groups, which explain the survival of the subjects. Essentially, our goal is to fit a pan-cancer model across multiple types of cancer. We wish to identify if there is any special characteristic which is common across these cancers. That way, we also learn about the presence of any important cancer-specific characteristic. Furthermore, by borrowing information among different cancers, we have a better power to identify important cancer-specific features.

Toward this goal, we consider r groups of cancers (or tumors) present in the data. Suppose that the interest is to make inference on the survival times of subjects of each cancer from p proteins which are same across all tumor groups. We use the AFT model, which regresses the survival time on the covariates. The AFT model is given by $\log(t_{ik}) = \sum_{j=1}^{p} x_{ijk}\,\beta_{jk} + \epsilon_{ik}$, $i = 1, \ldots, n_i$, $j = 1, \ldots, p, k = 1, \ldots, r$, where i denotes the patient, j denotes the protein, and k denotes the type of cancer. Likewise, t_{ik} is the survival time of the ith patient who has the kth cancer, x_{ijk} is the corresponding pth protein expression in the TCPA data, $\beta = (\beta_{11}, \ldots, \beta_{pr})$ is the vector of regression coefficients, and ϵ is the error vector. Assumption of $\epsilon \sim N(0, \sigma^2 I)$ gives rise to the log-normal AFT model, whereas, one could assume other distributions such as t distribution (Kleinbaum and Klein, 2006).

Letting c_{ik} be the censoring time, the observed time may be denoted by $t_{ik}^* = \min(t_{ik}, c_{ik})$; the corresponding observed censored indicator is $\delta_{ik} = I\{t_{ik} \le c_{ik}\}, I\{\cdot\}$ being the censoring indicator. Since the response is right censored, we follow the data augmentation approach of Tanner and Wong (1984) to impute the censored data w_{ik} (see also Bonato *et al.*, 2011), $w_{ik} = \log t_{ik}^*$, if t_{ik} is event time; and $w_{ik} > \log t_{ik}^*$, if t_{ik} is right censored.

2.1 | Shrinkage prior

Due to the presence of a large number of proteins, we will carry out a variable selection procedure in this AFT model to identify the important ones. We consider the shrinkage priors on the coefficients. In the shrinkage framework, a scale-mixture representation of the global-local priors allows parameters to be updated in blocks via a fairly automatic Gibbs sampler (Bhattacharya *et al.*, 2016) which makes it convenient for large-scale problems. Here, we adopt Horseshoe prior (Carvalho *et al.*, 2010; Peltola *et al.*, 2014). The hierarchical horseshoe representation for the AFT model is

$$\begin{gathered}\log t_{ik} \mid \beta_{jk}, \sigma^2 \sim N\left(\sum_{j=1}^{p} x_{ijk}\beta_{jk}, \sigma^2\right)\\ \beta_{jk} \mid \lambda_{jk}, \tau, \sigma^2 \sim N\left(b_{Pj}, \lambda_{jk}^2\tau^2\sigma^2\right), \quad b_{Pj} \sim N\left(0, \sigma_P^2\right)\\ \lambda_{jk} \sim C^+(0,1), \quad \tau \sim C^+(0,1),\\ \sigma^2 \sim \pi(\sigma^2) = 1/\sigma^2\end{gathered} \tag{1}$$

where $C^+(0, 1)$ is the truncated Cauchy density given by $f(x) = 1/\pi(1 + x^2), x > 0$. The conditional distribution and posterior computations are discussed in Web Appendix A.

2.2 | Borrowing strength

In the above formulation, a common τ allows borrowing strength across different cancer groups and proteins, while λ_{jk}'s provide protein and cancer-specific deviations. In addition, the different tumor groups do not have the same number of observations; hence straightforward regression for each tumor may not reveal the true picture of dependencies between the response and predictors, which will be further illustrated in Sections 4 and 5. In order to resolve this issue, we analyze the whole data together, that is, we integrate the data in our regression analysis. There are many advantages to doing so. For instance, it aids identifying nonobvious relationships existing in the data (Jansen *et al.*, 2002). Moreover, the interpretation of the result will be simultaneous and hence more easily understandable. In addition, individual regression for each tumor group may suffer from the lack of a sufficient number of observations, which we overcome through this type of data integration. The integration of the data can be done in several ways and Bayesian hierarchical models are particularly suitable for this purpose.

In order to take advantage of Bayesian hierarchical models, a convenient way is to borrow strength via the prior elicitation and carrying out joint estimation of the parameters across tumor groups. Ibrahim *et al.* (2002) specified a class of hierarchical priors on the regression coefficients in such a way that the correlation among covariates (proteins in our case) can be captured and strength can be borrowed across covariates. However, in our scenario, we wish to borrow strength across tumor groups, which can be accomplished by specifying a mean parameter for the coefficients in Equation (1).

The hyper-prior elicitation in Equation (1) helps us to create correlation among tumors for a given protein. The correlation between the kth group and the k'th group can be derived (Web Appendix G) as follows:

$$\mathrm{Corr}(\beta_{jk}, \beta_{jk'}) = \frac{\sigma_P^2}{\left(\lambda_{jk}^2\tau^2\sigma^2 + \sigma_P^2\right)^{\frac{1}{2}}\left(\lambda_{jk'}^2\tau^2\sigma^2 + \sigma_P^2\right)^{\frac{1}{2}}}. \tag{2}$$

We note that the correlation formulation depends on the hyperparameters σ_P, λ, τ, and σ. Since τ and σ are global and are not group-specific, they appear not to control borrow strength across tumor groups. In contrast, σ_P^2 is present in both the numerator and the denominator of the right-hand side in Equation (2), so increase in the value of σ_P^2 precludes increase in the value of correlation among proteins for different tumor groups and this is further illustrated in Web Appendix C. Nonetheless, it is the hyperparameter λ which helps us to borrow strength among tumors. Lower the value of λ_{jk}^2, higher will be the induced correlation. Note that since λ_{jk}^2 also serves as a shrinkage factor for the corresponding jth protein for the kth group, a lower value of λ_{jk}^2 will lead to declareg the protein as nonsignificant. Hence, although counter-intuitive, the maximum borrowing strength among tumor groups actually happens via nonsignificant proteins. This phenomenon will be observed via the posterior analysis of the correlations in TCPA data in Section 5.

Note that, within each group, an individual independent analysis can be carried out simply by setting the $b_{P_j} = 0$ in (1) which will be referred to as the "local" method in this paper.

3 | POSTERIOR INFERENCE FROM MCMC OUTPUT

In the Bayesian settings, the inference is based on the posterior samples of the MCMC chains. The survival curve prediction is outlined in the Web Appendix B. Other inferences proceed as follows.

To assess the variable selection performance in our simulation experiments, we consider the true-positive rate (TPR) or sensitivity, which is defined by the proportion of truly identified variables by an analysis which was originally present in the model. In contrast, the false-positive rate (FPR) is given by the proportion of falsely identified variables, which were not present originally. Plotting the TPR against FPR for varying cutoffs provides the so-called ratio operating characteristic (ROC) curve. It may be noted that, under a perfect fit, TPR should be close to 1 and FPR should be close to 0. This follows that the greater the area under the curve, the better is the fit.

The strategies described in the previous discussion provide convincing ways to compare different methodologies in simulation studies. Unfortunately, in TCPA data, we are unaware of the true proteins that are responsible for explaining the survival of the patients; rather our job is to exactly to find these to communicate with our clinical collaborators further. Due to the absence of a comprehensive path to do the variable selection, we propose a rank-based approach via a goodness-of-fit measure. To achieve this we use integrated Brier score (IBS; Hothorn *et al.*, 2006; Bonato *et al.*, 2011; the definition is provided in Web Appendix J).

We propose the following protein-selection scheme based on the IBS. The proteins are ordered according to their absolute value of the posterior means of the regression parameter estimates. Now a subset of topmost significant proteins can be used to fit an AFT model and the corresponding IBS can be calculated. Like forward selection criteria, if we keep increasing the number of proteins in the model per their order, the IBS will keep decreasing. Then the selection of ordered proteins proceeds until a desired value of IBS. A practical demonstration is provided in Section 5. We note that changing the threshold of IBS is equivalent to changing the threshold value to compute the TPR and FPR when we know the true model. In this way, depending on the threshold, we can recover the significant proteins that are common for more than one cancer group if they are included using the IBS threshold. For instance, in the example of tumors in the female body (see Web Appendix F), the protein DIRAS3 becomes significant for at least two cancer types, namely breast cancer and ovarian cancer.

4 | SIMULATION EXAMPLES

In this section, we discuss simulation scenarios to illustrate the methodologies discussed above. For the simulation, we consider five groups, each having 80 proteins. The sample sizes of the groups are set to 40, 50, 70, 100, and 120 respectively, so that, the first three groups have $p > n$ layouts. Furthermore, the total number of parameters (400) is more than the total number of observations (380). The covariates are generated from standard normal distribution independently. For the response generation, the σ^2 is set to 1. The censoring distribution is assumed to follow a Gamma distribution. In this way, the censoring rate can be varied generally by changing the shape and scale parameters accordingly. We simulate 100 data sets in this manner.

We consider two primary settings. In the first, which we refer to as Example 1, we generate β in the following manner. We assume the first protein is significant for all groups. Then the second protein is significant for the first group only. Similarly, the third, fourth, fifth, and sixth proteins are significant for the second, third, fourth, and fifth groups only, respectively. Then we keep replicating this generation of β for every other 10 variables. In order to make a protein significant, we set the corresponding $\beta = 1$. So, for each block of 10 proteins, the data generating matrix for β will look like:

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

after which, we assign a random positive or negative sign on each element.

We fit five different models in the generated data. Independent AFT regression with horseshoe prior is fitted for each group. This is referred to as "local." Then our proposed model discussed in Section 2.2 is fitted, which can be referred as "group-corr." As indicated in Web Appendix H, we also fit the model that borrows strength across groups and across proteins simultaneously ("all-corr"). In addition to these Bayesian methods, we apply two penalized cox regressions—lasso (Tibshirani, 1997; Simon *et al.*, 2011) and adaptive lasso, "alasso" (Li *et al.*, 2015)—for each group separately.

In the setting of this example, we declare a variable as significant when the absolute estimated β is more than a given threshold value. Once the variables are identified using a given threshold, we can compute the TPR and FPR. For 100 simulations of this experiment, the mean of these 100 TPRs and FPRs can be calculated. Now varying the threshold from a range between 0 and 1 will produce a series of mean TPRs and FPRs, which can be used to plot the ROC curve, given in Figure 1, for each method. One can note that the area under the ROC curves for the borrowing strength structure is greater than the that due to other methods, which indicates that whatever threshold is chosen, the variable selection performance of our proposed methodology remains superior. The area under the ROC curve (AUC) results are given in Table 2, which confirm our findings. For the Bayesian analyses, the posterior means have been calculated using 10 000 samples after 5000 burnin. From the ROC plots and the numerical results, it is evident that correlation structures actually help to estimate the true parameters.

In Example 2, we generate the simulated data under a similar set up as in Example 1, however, we consider a correlation structure among groups. In what follows, we produce the design matrix in a manner such that the rows of two different tumor groups are highly correlated with correlation 0.8. We plot the ROC curves in Figure 2 and report the AUC values for these ROC's in Table 3. Table 3 exhibits consistent performances of all methods as in the previous example.

5 | TCPA PROTEIN EXPRESSION DATA

We apply the methodologies described above in two types of TCPA data—kidney tumors and tumors in female bodies. While the first is described below, the second application is discussed in the Web Appendix F.

There are three types of kidney cancers viz. kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), and kidney renal papillary cell carcinoma (KIRP). According to Linehan and Ricketts (2013), all types of kidney cancers are different, making it even more important to characterize each one. In 2017, it is estimated that there were 63 990 new cases of kidney cancer and 14 400 deaths as a result of this disease (American Cancer Society 2017 report). Chromophobe (KICH) kidney cancer accounts for 5% of these cancer cases. In contrast, renal cell carcinoma is the most common type of kidney cancer, which is broadly classified into KIRC and KIRP.

The protein data for these tumors have 63, 469, and 215 samples, respectively, with 189 proteins, which follows that KICH group has a greater number of proteins than the number of observations whereas the KIRP consists of nearly the same number of samples and proteins. In addition, approximately 75.9% of samples are censored and Web Figure S1 presents group-wise observed Kaplan-Meir plots.

For finding the significant protein selections associated with the survival cancer outcome, these data are

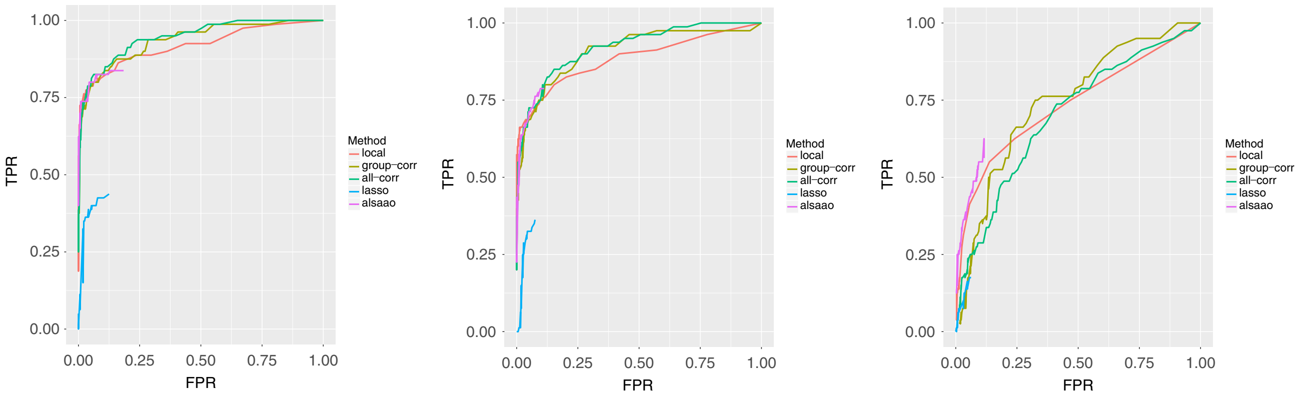

FIGURE 1 The ROC curves for different methods, from left to right subplots are for varying censoring rates—around 35%, 50%, and 76%, respectively. FPR, false-positive rate; ROC, ratio operating characteristic; TPR, true-positive rate [This figure appears in color in the electronic version of this article, and any mention of color refers to that version]

TABLE 2 Area under the ratio operating characteristic curves plotted in Figure 1

Censoring rate	35%	48%	76%
local	0.659	0.646	0.590
group-corr	0.695	0.689	0.648
all-corr	0.694	0.689	0.648
lasso	0.415	0.406	0.452
alasso	0.632	0.621	0.406

TABLE 3 Area under the ratio operating characteristic curves plotted in Figure 2, correlated scenario

Censoring rate	35%	56%	76%
local	0.661	0.649	0.597
group-corr	0.712	0.709	0.660
all-corr	0.712	0.708	0.655
lasso	0.377	0.444	0.460
alasso	0.557	0.388	0.418

used to apply our methodologies developed in this paper. In Web Figure S3 we plot the posterior estimates of the coefficients for a single MCMC chain. We insert the posterior summary for the group-corr method along with summary for the regressions run within each tumor group (local). All proteins have been standardized before the analyses. For the statistical inference, 20 000 posterior samples are collected after discarding 10 000 burnin samples.

5.1 | Protein selection

One of the key goals of the clinical researchers is to predict the survival times of the individuals, which can be accomplished by computing the posterior predictive survival curve. In addition, as mentioned in Section 3, we aim to select important proteins using this predictive summary. The mean IBS (defined in Section 3) obtained from four MCMC chains for group-corr method is 0.168 while the mean IBS for the local method is 0.201, which implies that the group-corr method has better predictive ability. It should be mentioned here that IBS due to all-corr method is same (0.168) as that due to the group-corr method. So, for simplicity, we restrict ourselves considering the group-corr method only for the following discussion.

In particular, parsimonious model selection is desirable in the presence of many features and hence sparsity is assumed in these kinds of analyses because of the easy interpretation. Toward this end, we identify the top 8 most significant proteins from the posterior estimates of the group-corr method and use these proteins to run a simple AFT model with log-normal assumption. This gives us the IBS as 0.172, which is very close to the IBS of original fitting with all proteins. So, one can conclude in favor of these eight proteins explaining maximum variation present in the data. Again, to confirm our outcome, we run a similar AFT model 10 times with randomly selected eight proteins, which results in IBS as 0.331, which is more than the IBS 0.172 due to group-corr. Since models with lower IBS are preferred, the variable selection technique using the correlation can be considered a reliable technique. The selected proteins are listed in Table 4. The effects of these proteins on the cancers have been well studied in the literature. For example, Advani *et al.* (2015) showed how medication to CRAF_pS338 improves treatment; Duckworth *et al.* (2016) concluded that overexpression of GAB2 promotes tumor growth; similarly, SF2 has been established as a critical pathway for human cancer cell survival, dissemination, and resistance to drug therapy (Wang *et al.*, 2014). PCADHERIN, FOXO3A_pS318S321, and DIRAS3

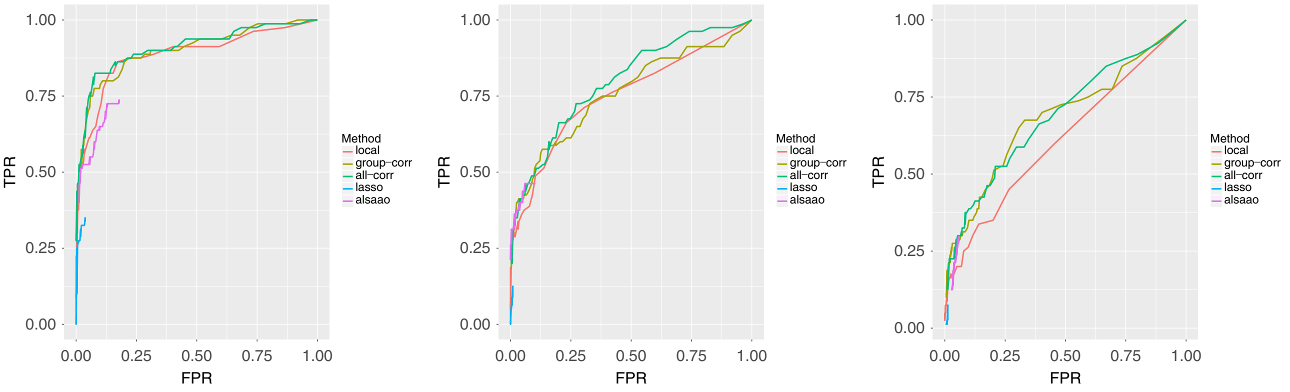

FIGURE 2 The ROC curves for different methods, from left to right subplots are for varying censoring rates—around 35%, 56%, and 76%, respectively, in the presence of correlation. FPR, false-positive rate; ROC, ratio operating characteristic; TPR, true-positive rate [This figure appears in color in the electronic version of this article, and any mention of color refers to that version]

TABLE 4 Top 8 proteins selected for kidney cancers

Chain 1	Chain 2	Chain 3	Chain 4
PCADHERIN	PCADHERIN	PCADHERIN	PCADHERIN
FOXO3A_pS318S321	DIRAS3	FOXO3A_pS318S321	DIRAS3
DIRAS3	FOXO3A_pS318S321	DIRAS3	FOXO3A_pS318S321
SF2	RAD51	RAD51	RAD51
RAD51	SF2	GAB2	SF2
GAB2	GAB2	SF2	GAB2
HER3_pY1298	BETACATENIN	HER3_pY1298	CRAF_pS338
CRAF_pS338	HER3_pY1298	BAK	HER3_pY1298

are the top 3 proteins recovered by all four chains. Not surprisingly, these are well-known for kidney tumor growth and invasion (Blaschke *et al.*, 2002; Ni *et al.*, 2014; Chen *et al.*, 2016).

6 | CONCLUSION

In this paper, we have proposed a Bayesian variable selection technique which accommodates both high-dimensional shrinkage and integration of the censored data. We have only considered the log-normal AFT model set up. Nonetheless, the extension to any other distribution is immediate. Furthermore, the use of the latent variable technique may also make possible extending this method for high-dimensional Bayesian Cox regression. Another future research topic will be to explore borrowing strength from multiple factors efficiently. For instance, Kling *et al.* (2015) considered the incorporation of sample sizes with the hope of eliminating the sample size effects of different groups from the final inference. One could follow their suggestion to incorporate similar factor terms in the prior elicitation.

Instead of using an improper prior, a vague Inverse Gamma prior on σ^2 may be more suitable. Another recommendation could be to integrate out parameters in obtaining the marginal distributions of w_{ik} which will be a Student's t distribution by suitably choosing the shape and rate parameters of inverse gamma prior on σ^2 (Sha *et al.*, 2006). Other suggestions include placing a hyper-prior on σ_P^2. For example, Ibrahim *et al.* (2002) argued for specifying an inverse gamma prior would help to borrow strength more. In contrast, Polson and Scott (2012) recommended a half-Cauchy prior. Nevertheless, in our set-up, we did not see much improvement in the estimations using these priors.

Future research is due to explore the impact of different sample sizes of the tumor groups. Moreover, the proposed method has been applied solely on two real-world data sets, and thus the effectiveness of the method should be thoroughly tested in the future on further applications. Nonetheless, our contribution toward the overreaching goal of extracting broad information out of the RPPA-based data bank TCPA will help to gain a better understanding of pan-cancer models.

ACKNOWLEDGMENTS

Maity, Mallick, and Baladandayuthapani were supported by National Institutes of Health under award number R01CA194391. Bhattacharya was supported by National Science Foundation Career award DMS 1653404. Baladandayuthapani was supported by National Institutes of Health under award numbers R01CA160736, R21CA220299, P30CA46592, National Science Foundation under award number DMS1463233, and start-up funds from the University of Michigan Rogel Cancer Center and School of Public Health. The authors thank the editor, the associate editors, and the reviewers for their helpful suggestions which substantially improved this article.

ORCID

Arnab K. Maity http://orcid.org/0000-0002-6692-0155

REFERENCES

Advani, S.J., Camargo, M.F., Seguin, L., Mielgo, A., Anand, S., Hicks, A.M. *et al.* (2015) Kinase-independent role for CRAF-driving tumor radioresistance via CHK2. *Nature Communications*, 6, 6.

Akbani, R., Ng, P.K.S., Werner, H.M., Shahmoradgoli, M., Zhang, F. Ju, Z. *et al.* (2014) A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature Communications*, 5, 3887.

Baladandayuthapani, V., Talluri, R., Ji, Y., Coombes, K.R., Lu, Y., Hennessy, B.T. *et al.* (2014) Bayesian sparse graphical models for classification with application to protein expression data. *The Annals of Applied Statistics*, 8(3), 1443.

Bhattacharya, A., Chakraborty, A. and Mallick, B.K. (2016) Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4), 985–991.

Bhattacharya, A., Pati, D., Pillai, N.S. and Dunson, D.B. (2015) Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512), 1479–1490.

Blaschke, S., Mueller, C.A., Markovic-Lipkovski, J., Puch, S., Miosge, N., Becker, V. *et al.* (2002) Expression of cadherin-8 in renal cell carcinoma and fetal kidney. *International Journal of Cancer*, 101(4), 327–334.

Bonato, V., Baladandayuthapani, V., Broom, B.M., Sulman, E.P., Aldape, K.D. and Do, K.-A. (2011) Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3), 359–367.

Brier, G. (1950) Verification of forecasts expressed in term of probabilities. *Monthly Weather Review*, 78, 1–3.

Cai, T., Huang, J. and Tian, L. (2009) Regularized estimation for the accelerated failure time model. *Biometrics*, 65(2), 394–404.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010) The horseshoe estimator for sparse signals. *Biometrika*, 97(2), 465–480.

Chen, W., Hill, H., Christie, A., Kim, M.S., Holloman, E., Pavia-Jimenez, A. *et al.* (2016) Targeting renal cell carcinoma with a HIF-2 antagonist. *Nature*, 539(7627), 112.

Cox, D.R. (1972) Regression models and life-tables. *Journal of the Royal Statistical Society, Series B (Methodological)*, 34(2), 187–220.

Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J.A., Sempoux, C. *et al.* (2009) A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, 1(4), 39.

Duckworth, C., Zhang, L., Carroll, S., Ethier, S. and Cheung, H. (2016) Overexpression of GAB2 in ovarian cancer cells promotes tumor growth and angiogenesis by upregulating chemokine expression. *Oncogene*, 35(31), 4036–4047.

George, E.I. and McCulloch, R.E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423), 881–889.

Hamid, J.S., Hu, P., Roslin, N.M., Ling, V., Greenwood, C.M. and Beyene, J. (2009) Data integration in genetics and genomics: methods and challenges. *Human Genomics and Proteomics*, 1(1), 1–13.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and Van Der Laan, M.J. (2006) Survival ensembles. *Biostatistics*, 7(3), 355–373.

Huang, J. and Ma, S. (2010) Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*, 16(2), 176–195.

Huang, J., Ma, S. and Xie, H. (2006) Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*, 62(3), 813–820.

Ibrahim, J.G., Chen, M.-H. and Gray, R.J. (2002) Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association*, 97(457), 88–99.

Jansen, R., Lan, N., Qian, J. and Gerstein, M. (2002) Integration of genomic datasets to predict protein complexes in yeast. *Journal of Structural and Functional Genomics*, 2(2), 71–81.

Khan, M.H.R. and Shaw, J.E.H. (2016) Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing*, 26(3), 725–741.

Khan, M.H.R. and Shaw, J.E.H. (2017) Variable selection for accelerated lifetime models with synthesized estimation techniques. *Statistical Methods in Medical Research*, 1–17.

Kleinbaum, D.G. and Klein, M. (2006) *Survival Analysis: A Self-Learning Text*. New York, NY: Springer Science & Business Media.

Kling, T., Johansson, P., Sanchez, J., Marinescu, V.D., Jörnsten, R. and Nelander, S. (2015) Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic Acids Research*, 43(15), 98–98.

Lee, K.E. and Mallick, B.K. (2004) Bayesian methods for variable selection in survival models with application to DNA microarray data. *Sankhyā: The Indian Journal of Statistics*, 66(4), 756–778.

Li, H. and Luan, Y. (2002) Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, 8, 65.

Li, J., Akbani, R., Zhao, W., Lu, Y., Weinstein, J.N., Mills, G.B. *et al.* (2017) Explore, visualize, and analyze functional cancer proteomic data using the Cancer Proteome Atlas. *Cancer Research*, 77(21), e51–e54.

Li, J., Lu, Y., Akbani, R., Ju, Z., Roebuck, P.L., Liu, W. *et al.* (2013) TCPA: a resource for cancer functional proteomics data. *Nature Methods*, 10(11), 1046–1047.

Li, X., Zeng, D. and Wang, Y. (2015) Coxnet: Regularized Cox Model. R Package V0.2.

Li, Y., Wang, J., Ye, J. and Reddy, C.K. (2016) A multi-task learning formulation for survival analysis. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 1715-1724.

Linehan, W.M. and Ricketts, C.J. (2013) The metabolic basis of kidney cancer. *Seminars in Cancer Biology*, 23(1), 46–55.

Liu, B., Li, Y., Sun, Z., Ghosh, S. and Ng, K. (2018, February) *Early Prediction of Diabetes Complications from Electronic Health Records: A Multi-Task Survival Analysis Approach*. Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA.

Miller, R.G. (1976) Least squares regression with censored data. *Biometrika*, 63(3), 449–464.

Narisetty, N.N. and He, X. (2014) Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2), 789–817.

Ni, D., Ma, X., Li, H.-Z., Gao, Y., Li, X.-T., Zhang, Y. *et al.* (2014) Downregulation of FOXO3a promotes tumor metastasis and is associated with metastasis-free survival of patients with clear cell renal cell carcinoma. *Clinical Cancer Research*, 20(7), 1779–1790.

Park, E.S., Rabinovsky, R., Carey, M., Hennessy, B.T., Agarwal, R., Liu, W. *et al.* (2010) Integrative analysis of proteomic signatures, mutations, and drug responsiveness in the NCI 60 cancer cell line set. *Molecular Cancer Therapeutics*, 9(2), 257–267.

Peltola, T., Havulinna, A.S., Salomaa, V. and Vehtari, A. (2014) Hierarchical Bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop*, CEUR-WS.org, volume 1218, pp. 79-88.

Polson, N.G. and Scott, J.G. (2012) On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4), 887–902.

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

Shankavaram, U.T., Reinhold, W.C., Nishizuka, S., Major, S., Morita, D., Chary, K.K. *et al.* (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Molecular Cancer Therapeutics*, 6(3), 820–832.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2011) Regularization paths for Coxs proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5), 1–13.

Tanner, M.A. and Wong, W.H. (1984) Data-based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis. *Journal of the American Statistical Association*, 79(385), 174–182.

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

Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4), 385–395.

Wang, J., Su, L., Chen, X., Li, P., Cai, Q., Yu, B. *et al.* (2014) Malat1 promotes cell proliferation in gastric cancer by recruiting sf2/asf. *Biomedicine & Pharmacotherapy*, 68(5), 557–564.

Wang, L., Li, Y., Zhou, J., Zhu, D. and Ye, J. (2017) Multi-task survival analysis. *2017 IEEE International Conference on Data Mining (ICDM)*, IEEE, pp. 485-494.

Wang, X. and Song, L. (2011) Adaptive Lasso variable selection for the accelerated failure models. *Communications in Statistics—Theory and Methods*, 40(24), 4372–4386.

Wei, L.-J. (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15), 1871–1879.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K. *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10), 1113–1120.

Zhang, Z., Sinha, S., Maiti, T. and Shipp, E. (2018) Bayesian variable selection in the AFT model with an application to the SEER breast cancer data. *Statistical Methods in Medical Research*, 27(4), 971–990.

Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

SUPPORTING INFORMATION

Web Appendix A referenced in Section 2.1, Web Appendix B referenced in Section 3, Web Appendix C referenced in Section 2.2, Web Appendix D referenced in Sections 3 and 5, Web Appendix G referenced in Section 2.2, Web Appendix H referenced in Section 4, Web Appendix J referenced in Section 3, and Web Figure S1 referenced in Section 5 are available with this paper at the Biometrics Website on Wiley Online Library. The R package PanCanVarSel implementing the posterior MCMC is available on R CRAN.

How to cite this article: Maity AK, Bhattacharya A, Mallick BK, Baladandayuthapani V. Bayesian data integration and variable selection for pan-cancer survival prediction using protein expression data. *Biometrics*. 2020;76:316–325. https://doi.org/10.1111/biom.13132