Bioinformatics, 36(13), 2020, 3951–3958
doi: 10.1093/bioinformatics/btaa286
Advance Access Publication Date: 5 May 2020
Review

Gene expression

Bayesian structural equation modeling in multiple omics data with application to circadian genes

Arnab Kumar Maity[1,*], Sang Chan Lee[2], Bani K. Mallick[2] and Tapasree Roy Sarkar[2,3]

[1]Early Clinical Development Oncology Statistics, Pfizer Inc., San Diego, CA 92121, USA, [2]Department of Statistics, and [3]Department of Biology, Texas A&M University, College Station, TX 77843, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on October 18, 2019; revised on March 30, 2020; editorial decision on April 23, 2020; accepted on April 27, 2020

Abstract

Motivation: It is well known that the integration among different data-sources is reliable because of its potential of unveiling new functionalities of the genomic expressions, which might be dormant in a single-source analysis. Moreover, different studies have justified the more powerful analyses of multi-platform data. Toward this, in this study, we consider the circadian genes' omics profile, such as copy number changes and RNA-sequence data along with their survival response. We develop a Bayesian structural equation modeling coupled with linear regressions and log normal accelerated failure-time regression to integrate the information between these two platforms to predict the survival of the subjects. We place conjugate priors on the regression parameters and derive the Gibbs sampler using the conditional distributions of them.

Results: Our extensive simulation study shows that the integrative model provides a better fit to the data than its closest competitor. The analyses of glioblastoma cancer data and the breast cancer data from TCGA, the largest genomics and transcriptomics database, support our findings.

Availability and implementation: The developed method is wrapped in R package available at https://github.com/MAITYA02/semmcmc.

Contact: arnab.maity@pfizer.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In the current era of precision medicine, each subject is targeted for treatment modeled via individual healthcare data. To this end of advanced treatment, it is of interest the molecular profiling besides the clinical profiling of the patients. Accurate prognostic prediction using molecular profiles is an essential ingredient to develop precision medicine. Under this regime, cancer studies that are focused on 1D omics data have only provided limited information regarding the etiology of oncogenesis and tumor progression (Huang *et al.*, 2017). To overcome this problem, scientists have focused to integrate multi-platform data in cancer research.

The advent of multi-platform data has been directing the biological research and statistical methodological research to collect and analyze these multi-platform data. The Cancer Genome Atlas (TCGA) is the largest collection of parallel transcriptomics, genomics and proteomics data along with patient's demographic information, primary aim of which is to generate, quality control, merge, analyze and interpret molecular profiles at the DNA, RNA, protein and epigenetic levels for hundreds of clinical tumors representing various tumor types and their subtypes (Weinstein *et al.*, 2013). Cases that meet quality assurance specifications are characterized using technologies that assess the sequence of the exome, copy number variation (CNV, measured by SNP-arrays), DNA methylation, mRNA expression and RNA sequence, microRNA (miRNA) expression and transcript splice variation, whole-genome sequencing and reverse phase protein arrays. Attention is being paid to identify the genomic alterations across these platforms to improve the therapeutic response, which may be evident from the phenotypical measures, such as survival of the cancer patients. The reasoning behind this attention can be motivated by each of the hundreds of genetic alterations inside of a genome providing a complementary view of the underlying complex biological process and thus an integrative analysis of multiple platform is required to achieve the overreaching goal of cancer studies.

Circadian oscillation is a fundamental process that regulates a wide variety of physiological and metabolic processes. Perturbations of circadian rhythmicity are associated with significant physiological consequences including metabolic disorders and cancer (Sahar and Sassone-Corsi, 2009). Increased cancer incidence and progression have often been linked to disruption or deregulation of the molecular mechanism of the circadian clock (Fu and Kettner, 2013). Circadian rhythms are referred to those organisms, which exhibit time dependent behavior across a 24-h day. These outputs are driven

by manifestations of phasic cyclic gene expression patterns. Nearly half of all protein-coding genes show circadian-dependent transcription in at least one tissue in mammals (Andreani *et al.*, 2015). There is increasing evidence that links dysfunction of the clockwork with the pathogenesis of cancer, such as breast cancer and brain cancer (Davis and Mirick, 2006). In this article, we propose a Bayesian structural equation coupled with Bayesian accelerated failure time (AFT) model to carry out an integrative analysis where the integration takes place among the multiple platform of omics data. We consider some important circadian genes, which have been reported to play an important role in breast and brain cancer progression.

We note that, the direction of biological relationship is arbitrary and it may be a good practice to introduce some latent variables along with the observed variables to describe the relationships. To this end, Wong *et al.* (2018) proposed structural equation modeling (SEM) to model the TCGA data. The history of SEM dates back to Bentler and Weeks (1980) and it has been used extensively in the literature thereafter; e.g. in psychology (Quintana and Maxwell, 1999), in economics (Heckman and Vytlacil, 2005) and in healthcare sector (Naliboff *et al.*, 2012). SEM requires the introduction of latent variables and there are several studies, which make use of latent variable for survival regressions, e.g. in Cox proportional-hazard model (Larsen, 2005; Stoolmiller and Snyder, 2006).

The concept of integration is very broad. Generally, based on the direction, they can be classified into two broad groups—horizontal and vertical (Chu and Huang, 2017). In the horizontal integration analysis, omics data of same types but different studies or laboratories are combined. On the other hand, when the different omics data for the same patient is analyzed then it is called the vertical integration, which is the focus of this study. The vertical integration methods are then categorized into different groups depending on the methodologies used. For instance, Bayesian and non-Bayesian integration methods, network-based integration method, supervised learning and non-supervised learning etc. For a full review, we refer the readers to Huang *et al.* (2017). Other comprehensive references include Tseng *et al.* (2015), Gomez-Cabrero *et al.* (2014) and Hamid *et al.* (2009). A popular network-based method was reported by Vaske *et al.* (2010), who developed a supervised graphical model incorporating the pathway information. Another example of unsupervised learning is iCluster method by Shen *et al.* (2009), where by using the penalized likelihood approach they derived a clustering solution for tumor cells. Daemen *et al.* (2009) proposed a kernel-based support-vector machine to integrate microarray and omics data for the cancer patients. However, many of these methods do not consider the underlying biological relationship between multiple omics data-sources.

As a remedy, Wang *et al.* (2013) proposed an integrated Bayesian model, which essentially combines a two-stage regressions in a unified manner. The first model regresses the gene expression on the methylation expression, and the second model then regresses the clinical variable or the phenotype on the estimated effects from the first model. However, a major criticism of this model is that statistically it encourages increment of the errors when going from the first model to the second model.

Nevertheless, the Bayesian paradigm for structural equation is notably scant; important references include Palomo *et al.* (2007) and Song and Lee (2012). Among them, the work of Song and Lee (2012) has described the basic ingredients of Bayesian SEM with few examples and the codes are written in WinBUGS. However, to the best of our knowledge, there has been no study on the application of SEM under the Bayesian regime in survival settings. In this article, we propose a Bayesian structural equation coupled with Bayesian AFT model to carry out an integrative analysis where the integration takes place among the multiple platform of omics data. We consider the DNA CNV and RNAseq data-sources as the two platforms to predict the survival of the patients. We show that an integrative analysis outperforms the usual regression model where the underlying biological relationship is not captured.

In general, the TCGA collects and provides various levels DNA-level data—methylation expression, mutation and DNA copy number changes (Wang *et al.*, 2013). These molecular features coupled with miRNA expressions data are known to affect the gene expression-level data measured by microarray technology or by next-generation RNA-sequencing technology. The genes then code for proteins, which directly controls the tumor growth. This relationship is schematically displayed in Figure 1. Any integration method, which wants to consider the underlying direction among the platforms faces additional challenge and thus requires additional processing. For example, each transcriptomics factor may or may not be responsible for over-expression or under-expression for either one, or multiple or neither of the genomics features. In a very similar fashion, each gene expression may or may not code proteins and can affect the function of multiple proteins, which are the primary factors for tumor growth or tumor suppression. To overcome this, we assume that the expressions of each platform are controlled by a latent variable and the latent variables from other platforms. The details are described in Section 2 (Fig. 1).

In this study, we introduced the Bayesian methodologies of TCGA data using a structural equation model and used the posterior analysis via the Markov Chain Monte Carlo (MCMC). Our model formulation is similar in the spirit of what is proposed by Wong *et al.* (2018). However, they considered Cox proportional-hazard model in order to model the survival time and used the EM algorithm technique to maximize the likelihood function. In addition, they assumed the latent variable can be measured via the various types of gene expressions for a single gene and hence their model could consider a single gene at a time. In the contrary, we assume that for multiple genes, there exist a latent variable for circadian gene expressions or CNVs, so this can easily accommodate multiple genes in a single model for a better result. We considered glioblastoma cancer and breast cancer datasets for a set of genes, which have been known to affect the circadian rhythms. For integration among different platforms, we consider two types of measurements of those genes—copy number changes and normalized RNA-sequencing data. Our model showed that integration among these two platforms provides a better fit for the survival outcome of the subjects.

The remainder of this article is organized as follows. Section 2 introduces the Bayesian methodologies of TCGA data using a structural equation model and provides a brief description how to carry the posterior analysis via the MCMC. In Section 3, we describe simulation examples and show that the performance of our proposed model is superior to the general kind of regression. We then illustrate our methodology by applying to TCGA cancer data in Section 4. We consider two cancers, namely glioblastoma cancer and breast cancer datasets for a set of genes, which have been known to affect the circadian rhythms of a human biological clock. For integration among different platforms, we consider two types of measurements of those genes—copy number changes and normalized RNA-sequencing data. Using our method, we justify that integration among these two platforms provide a better fit for the

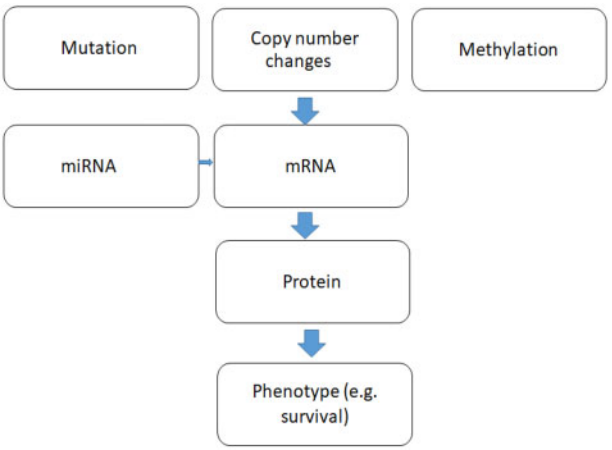

Fig. 1. Biological relationships among gene expressions data platforms

survival outcome of the subjects. The discussions and conclusions in Section 5 are then followed.

2 Multiple omics data-integrated model

2.1 The model

Vertical integration is referred to the analysis when the different data are collected from multiple transcriptomics, genomics and proteomics platforms for a same subject to infer about the cell outcomes. To ease of explanation, we provide the model development strategy for two platforms namely, CNV and gene expression (mRNA), however can be generalized for multiple platforms in a straightforward manner. The phenotypical model, we consider here is the log-normal AFT model for the survival outcome with demographic variables as the covariates.

In what follows, we assume that each platform gets affected and can be explained by a latent variable. Let n be the number of individuals, q_1 be the number of mRNA expressions and q_2 be the number of CNV measurements. Also let η_1 and η_2 be the latent variables, which control the mRNA expressions, $\mathrm{U}_{1k}, k = 1, \ldots, q_1$, and CNV measurements, $\mathrm{U}_{2l}, l = 1, \ldots, q_2$, respectively. Each gene measurement is for n individuals and hence a $n \times 1$ vector. In addition, η_1 can also be explained by η_2, meaning that the significance of the copy number changes are captured to describe the mRNA expressions. Finally, we construct the AFT regression model of survival data $(\mathrm{t}^*, \delta) = ((t_1^*, \delta_1), \ldots, (t_n^*, \delta_n))'$ with some covariates $\mathrm{x}_j = (x_{1j}, \ldots, x_{nj})', j = 1, \ldots, p$.

Here, δ_i is the censored indicator and takes 1 if a death is observed and takes 0 if right censored. Given the actual death time t and censoring time c are independent, $t_i^* = \min(t_i, c_i)$. Hence, we propose the following structural equation model:

$$\begin{aligned} \log\ t &= \alpha_t + \mathrm{X}\beta_t + \eta_1\phi_t + \epsilon_t \\ \mathrm{U}_{1k} &= \alpha_{u_1 k} + \eta_1\phi_{u_1 k} + \epsilon_{u_1 k}, \quad k = 1, \ldots, q_1, \end{aligned} \tag{1}$$

$$\mathrm{U}_{2l} = \alpha_{u_2 l} + \eta_2\phi_{u_2 l} + \epsilon_{u_2 l}, \quad l = 1, \ldots, q_2. \tag{2}$$

Here, ϵ_t is the error vector for the AFT regression model. Assuming $\epsilon_t \sim N(0, \sigma_t^2 \mathrm{I})$ gives raise to the log-normal AFT model. In addition, we assume $E(\epsilon_{u_1}) = E(\epsilon_{u_1}) = 0$ and $Cov(\epsilon_{u_1}, \epsilon_{u_2}) = 0$. $(\alpha_t, \alpha_{u_1}, \alpha_{u_2})$ are the intercept parameters and $(\beta_t, \phi_t, \phi_{u_1}, \phi_{u_2})$ are the suitable regression parameters. To carry out the analysis in Bayesian fashion, we impute the censored observations from the appropriate truncated normal distribution. In addition, we assume that $\epsilon_{u_1 k} \sim N(0, \sigma_{u_{1k}}^2 I)$ and $\epsilon_{u_2 l} \sim N(0, \sigma_{u_{2l}}^2 I)$ such that each of (1) and (2) is a standard linear regression model. Furthermore, while it is assumed that η_1 is dependent on η_2 via $\eta_1 \sim N(\eta_2, \sigma_{\eta_1}^2)$, η_2 assumed to be independently follow $N(0, \sigma_{\eta_2}^2)$. The schematic diagram of our structural equation model is shown in Figure 2.

We assume the standard multivariate normal distribution on the regression coefficients β, which can be made a vaguely informative by assuming a large variance component in the variance-covariance matrix. Nevertheless, other regression parameters $\alpha_t, \alpha_{u_1 k}, \alpha_{u_2 l}, \phi_{u_1 k}, \phi_{u_2 l}$ are all assumed to follow a normal distribution and can be made a vaguely informative. While, we assume a non-informative prior on σ_t^2, the other variance parameters are kept as fixed for our study. With these ingredients, the full Bayesian hierarchical representation is

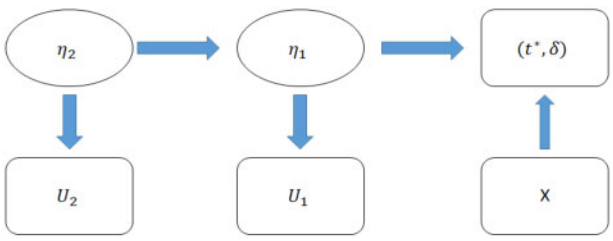

Fig. 2. Biological relationships among gene expressions data platforms

$$\log\ t \sim N(\alpha_t + \mathrm{X}\beta_t + \eta_1\phi_t, \sigma_t^2 \mathrm{I}), \tag{3}$$

$$\mathrm{U}_{1k} \sim N(\alpha_{u_{1k}} + \eta_1\phi_{u_{1k}}, \sigma_{u_1}^2 \mathrm{I}), \quad k = 1, \ldots, q_1, \tag{4}$$

$$\mathrm{U}_{2l} \sim N(\alpha_{u_{2l}} + \eta_2\phi_{u_{2l}}, \sigma_{u_2}^2 \mathrm{I}), \quad l = 1, \ldots, q_2, \tag{5}$$

$$\eta_1 \sim N(\eta_2, \sigma_{\eta_1}^2), \tag{6}$$

$$\begin{aligned} &\eta_2 \sim N(0, \sigma_{\eta_2}^2) \\ &\beta \sim N(\beta_{t0}, \sigma_t^2 \Sigma_\beta) \\ &\alpha_t \sim N(\alpha_{t0}, \sigma_t^2 \sigma_{\alpha_t}^2) \\ &\alpha_{u_{1k}} \sim N(\alpha_{u_{1k}0}, \sigma_{\alpha_{u_{1k}0}}^2), \quad k = 1, \ldots, q_1 \\ &\alpha_{u_{2l}} \sim N(\alpha_{u_{2l}0}, \sigma_{\alpha_{u_{2l}0}}^2), \quad l = 1, \ldots, q_2 \\ &\phi_t \sim N(\phi_{t0}, \sigma_\phi^2) \\ &\phi_{u_{1k}} \sim N(\phi_{u_{1k}0}, \sigma_{\phi_{u_{1k}0}}^2), \quad k = 1, \ldots, q_1 \\ &\phi_{u_{2l}} \sim N(\phi_{u_{2l}0}, \sigma_{\phi_{u_{2l}0}}^2), \quad l = 1, \ldots, q_2 \\ &\sigma_t^2 \sim \pi(\sigma_t) \equiv 1/\sigma_t^2. \end{aligned} \tag{7}$$

2.2 Identifiability

A common problem is identifiability of the full model when using a structural equation models. Bollen and Davis (2009) discussed few conditions under which a structural equation model becomes identifiable using the *exogenous X* rule. In what follows, we provide a brief description of the conditions and show that those hold under our formulation of the model.

First, each latent variable should have at least one observed variable that loads solely on it and the associated errors of measurement are uncorrelated. According to the formulation of our model, the observed matrices $\underline{u}_1$ and $\underline{u}_2$ are solely related to the latent variables η_1 and η_2, respectively. In addition, we have assumed that the corresponding error vectors are uncorrelated. So this suffices this condition. Second, each latent variable must have at least two observed indicators in total and the errors of these other indicators are uncorrelated with those of the unique indicators. This is satisfied trivially with the formulation of our model. Finally, the latent variable model (6) must have an identical structure, which is also true here.

2.3 Posterior computation

The posterior computation for the right censored data is not straightforward since the censored data are not originally observed. Nonetheless, we impute the right censored observations following the data augmentation approach (Bonato *et al.*, 2011; Tanner and Wong, 1987). We denote the augmented data vector by $\mathrm{y} = (y_1, \ldots, y_n)'$, where

$$\begin{cases} y_i = \log\ t_i^* & \text{if } \delta_i = 1 \\ y_i > \log\ t_i^* & \text{if } \delta_i = 0. \end{cases}$$

Hence, to carry out the posterior analysis, at the s-th iteration of the MCMC chain, the censored data are sampled from the truncated normal distribution

$$y_i^{(s)} \sim N(\alpha_t^{(s)} + \sum_{j=1}^{p} x_{ij}\beta_{tj}^{(s)}, \sigma_t^{2(s)}) I(y_i > \log\ t_i^*) \text{ if } \delta_i = 0.$$

In a similar fashion, with the latent variables in the joint likelihood, the posterior distribution becomes intractable. However, using the data augmentation scheme, the latent variables can be updated from the conditional distributions $\eta_1|. \sim N(\mu_{\eta_1 \text{post}}, \sigma_{\eta_1 \text{post}}^2)$ and $\eta_2|. \sim N(\mu_{\eta_2 \text{post}}, \sigma_{\eta_2 \text{post}}^2)$ respectively, where,

$$\sigma^{2}_{\eta_1\,\text{post}} = 1/\sigma^2_{\eta_1} + \Big(\sum_{k=1}^{q_1}\phi^2_{u_{1k}}/\sigma^2_{u_{1k}}\Big)$$

$$\mu_{\eta_1\,\text{post}} = 1/\sigma^{2}_{\eta_1\,\text{post}}\Big(\eta_2/\sigma^2_{\eta_1} + \Big(\sum_{k=1}^{q_1}\phi_{u_{1k}}\mathrm{u}_{1k}/\sigma^2_{u_{1k}}\Big) + \phi_t 1_n^T \log \mathrm{t}/\sigma^2_t - \sum_{k=1}^{q_1}\Big(\alpha_{u_{1k}}\phi_{u_{1k}}/\sigma^2_{u_{1k}}\Big) - \phi_t\alpha_t/\sigma^2_t - \beta_t^T \mathrm{X}^T\phi_t 1_n/\sigma^2_t\Big)$$

$$\sigma^{2}_{\eta_2\,\text{post}} = 1/\sigma^2_{\eta_1} + 1/\sigma^2_{\eta_2} + \sum_{l=1}^{q_2}\Big(\phi^2_{u_{2k}}/\sigma^2_{u_{2k}}\Big)$$

$$\mu_{\eta_2\,\text{post}} = 1/\sigma^{2}_{\eta_2\,\text{post}}\Big(\eta_1/\sigma^2_{\eta_1} + \sum_{l=1}^{q_2}\Big(\phi_{u_{2l}}u_{2l}/\sigma^2_{u_{2l}}\Big) + \sum_{l=1}^{q_2}\Big(\alpha_{u_{2l}}\phi_{u_{2l}}/\sigma^2_{u_{2l}}\Big).$$

Once the latent variables are updated they can be treated as observed as the other observed variables. Hence, the conditional distributions are available explicitly, and Gibbs sampling can be employed to cycle the iterations. The remaining details are as follows.

When, $\beta_{t0} = \alpha_{t0} = \phi_{t0} = 0$,

$$\begin{aligned}
\beta_t|. &\sim N(B^{-1}\mathrm{X}^T(\mathrm{y} - \alpha_t - \eta_1\phi_t), \sigma^2_t B^{-1}), \quad B = (\mathrm{X}^T\mathrm{X} + \Sigma_\beta^{-1})\\
\alpha_t|. &\sim N(A^{-1}1_n^T(\mathrm{y} - \mathrm{X}\beta_t - \eta_1\phi_t), \sigma^2_t A^{-1}), \quad A = (1^T1 + \sigma^2_{\alpha_t})\\
\alpha_t|. &\sim N(A^{-1}1_n^T(\mathrm{y} - \mathrm{X}\beta_t - \eta_1\phi_t), \sigma^2_t A^{-1}), \quad A = (1^T1 + \sigma^2_{\alpha_t})\\
\phi_t|. &\sim N(P^{-1}\eta_1 1_n^T(\mathrm{y} - \alpha - \mathrm{X}\beta_t), P^{-1}), \quad P = (\eta_1 1^T 1 + \sigma^2_{\phi_t})\\
\sigma^2_t|. &\sim \text{Inverse Gamma}[\text{shape} = (n + p + 1)/2,\\
\text{scale} &= \{(\mathrm{y} - \alpha_t - \mathrm{X}\beta_t - \eta_1\phi_t)^T(\mathrm{y} - \alpha_t - \mathrm{X}\beta_t - \eta_1\phi_t) + \beta_t^T\beta_t + \alpha_t^2\}/2].
\end{aligned}$$

Other parameters can be updated from similar conditional distributions by assuming η_1 and η_2 as observed once they are imputed in the MCMC chain.

2.4 Goodness of fit

There exist several model validation criteria, such as log pseudo marginal likelihood (LPML) (Gelfand *et al.*, 1992), L-measure (Ibrahim and Laud, 1994) or DIC (Spiegelhalter *et al.*, 2002). The literature is also advocated with the application of these criteria in survival settings, e.g. see Brown *et al.* (2005), Ibrahim *et al.* (2005) and Rizopoulos and Ghosh (2011). In this article, to measure the goodness of fit, we consider the deviance information criterion, which combines goodness of fit of a model with a penalty for model complexity and is defined as the model deviance + 2× (effective number of parameters), evaluated at a posterior point estimate of the parameter. In particular, $\text{DIC} = D(\bar{\theta}) + 2p_{\text{D}}$, where $D(\theta) = -2\log f(.|\theta)$, $f(.|\theta)$ is the likelihood function of the model and $\bar{\theta}$ is an estimate of the model parameter θ. In the above expression, p_{D} is termed as the effective number of parameters and is defined as $p_{\text{D}} = \text{D}(\bar{\theta}) - D(\bar{\theta})$, where $\text{D}(\bar{\theta})$ is a posterior point estimate of the deviance. In our proposed model, it is possible to partition the likelihoods over survival and coordinates and thus to obtain the DIC for survival model. A model with smaller value of DIC is preferred.

The conditional predictive ordinate (CPO) is a Bayesian model diagnostic criterion introduced in Geisser and Eddy (1979) and its implementation in sampling-based approaches is discussed in Gelfand *et al.* (1992). For a model, the CPO of the ith observation y_i is defined as

$$\text{CPO}_i = f(y_i|\mathrm{y}_{-i}) = \int f(y_i|\theta)\pi(\theta|\mathrm{y}_{-i})d\theta,$$

where $\mathrm{y}_{-i} = \mathrm{y} \setminus \{y_i\}$. Gelfand *et al.* (1992) provided an estimate of CPO_i based on Markov chain samples from the full posterior $\pi(\theta|\mathrm{y})$. The LPML of model $\ell(\alpha)$ is constructed similar to the log-likelihood, but based on the CPO_i, and is defined as $\text{LPML} = \log\prod_{i=1}^{n}\text{CPO}_i$. Model with higher LPML is preferred. The LPML is well defined, provided the predictive density is proper and thus may be defined under improper priors as well.

3 Operating characteristics in simulation studies

3.1 Integrated model as the data-generating model

In this section, we study some simulated examples to observe the prediction performance using our Bayesian structural equation integrated model. To this direction, we generate the covariate matrix from a multivariate normal distribution with mean 0, variance-covariance matrix as unit matrix and dimension two, i.e. p=5. All the regression parameters β and the intercept parameters $\alpha_t, \alpha_{u_1}, \alpha_{u_2}$ are generated from a uniform distribution $U(-1, 1)$. We set the latent variable coefficients $\phi_t = \phi_{u_1} = \phi_{u_2} = 1$ and the variance parameters $\sigma^2_t = 1$, $\sigma^2_{u_1} = \sigma^2_{u_2} = 1$. Additionally, we consider generating the data by setting $\sigma^2 = 2$ and $\sigma^2_{u_1} = \sigma^2_{u_2} = 1$. The case of varying $\sigma^2_{u_1}$ and $\sigma^2_{u_2}$ is discussed in Section 3.3, and the impact of placing an informative proper prior is discussed in the Supplementary Material.

Then the latent variables η_1 and η_2 are generated according to (6) and (7), respectively. The mRNA expressions $\underline{\mathrm{u}}_1$ and copy number changes $\underline{\mathrm{u}}_2$ are simulated according to (4) and (5), respectively. Finally, we simulate the time components t in log scale according to (3). We consider both situations with censoring and with no censoring. When censored subjects are created the censoring distribution of the censoring time c is assumed to follow a Gamma distribution and hence the amount of censored data can be controlled by varying the shape and scale parameters of the Gamma distribution. So, we obtain the observed paired response data $\{t^*_i, \delta_i\} = \{\min(t_i, c_i), I(t_i < c_i)\}, i = 1, \ldots, n$, where, $n = 100$. We simulate 100 similar datasets in order to assess the goodness of fit of the integrated model in repeated experiments.

When fitting the integrated model to the simulated data, we set all the mean parameters of the prior distributions as 0. In addition, the variance parameters of the normal priors are kept as 1, while the variance-covariance matrix for β_t is diag(100 000). We simulate the datasets with the censoring rates 0% (no censoring), 28%, 37% and 50%. It is observed that, in the Bayesian analysis after discarding 2000 burn-in samples, 100 000 iterations with 100 thinning provide a good stationary Markov chain. For comparison, we also fit a Bayesian log-normal AFT model on the data with covariates $\underline{\mathrm{x}}$, $\underline{\mathrm{u}}_1$ and $\underline{\mathrm{u}}_2$, i.e. the demographic variables, mRNA and CNV data, respectively, and referred to it as nonIntegrated model. In particular, we fit the model

$$\log\, t = \alpha + \mathrm{X}\beta + \mathrm{U}_1\gamma_1 + \mathrm{U}_2\gamma_2 + \epsilon, \tag{8}$$

where $\epsilon \sim N(0, \sigma^2 I)$, and $\beta, \gamma_1, \gamma_2$ are corresponding regression coefficients. We impose vaguely informative prior on the parameters as discussed previously and to carry out the Bayesian analysis we augment the censored data and impute them.

To the best of our knowledge, the existing software do not handle censored survival outcomes. Nevertheless, to compare with the available software packages, we selected the R package *lavaan* (Rosseel, 2012) as a representative. This method is used to estimate the MSE when there is no censoring in the data. A related comparison with the iBAG method (Wang *et al.*, 2013) is provided in the Supplementary Material.

The MCMC routine takes about 1.6 min per 100 000 iterations in a computing system equipped with Intel(R) Core(TM) i5-8350U CPU @ 1.76 GHz 1.90 GHz processor, 8.00 GB RAM and 64-bit operating system. Table 1 summarizes the result and the superior performance of the proposed integrated model is evident form the table. For instance, in the case of $\sigma^2_t = 1$, when about 28% data are right censored the DIC of integrated model is 290.04 while the same for the nonIntegrated model is 314.25; this suggests that the integrated method, where the underlying relationship is captured, provides a better fit to the data. Similarly, the LPML due to integrated method is −313.82 that is greater than −787.30, LPML due to the nonIntegrated method, which supports in favor of the structural equation model-based integration method.

Furthermore, the existing *lavaan* package employs a non-Bayesian method to fit the SEM and hence the model fitting criteria, such as DIC and LPML, cannot be computed. When the MSE is calculated for the non-censoring case, we notice that the estimated

Table 1. Goodness of fit for the integrated and nonIntegrated models in simulation examples

Censor rate	Method	DIC	LPML	MSE
$\sigma_t^2 = 1$				
0%	Integrated	276.32	−575.22	0.000
	NonIntegrated	284.42	−802.14	0.000
	lavaan	—	—	2.362
28%	Integrated	290.04	−313.82	0.048
	NonIntegrated	314.25	−787.30	0.052
37%	Integrated	288.07	−283.90	0.113
	NonIntegrated	308.30	−815.26	0.221
50%	Integrated	300.14	−235.73	0.296
	NonIntegrated	303.02	−1267.44	0.345
$\sigma_t^2 = 2$				
0%	Integrated	417.46	−344.67	0.000
	NonIntegrated	441.50	−972.82	0.000
	lavaan	—	—	4.154
28%	Integrated	395.24	−255.26	0.623
	NonIntegrated	409.13	−556.66	0.734
37%	Integrated	386.99	−197.14	0.946
	NonIntegrated	397.52	−467.39	1.095
50%	Integrated	356.88	−159.57	1.345
	NonIntegrated	376.53	−491.02	1.480
Data-generating model is the nonIntegrated model				
0%	Integrated	553.18	−497.72	0.000
	NonIntegrated	545.42	−441.32	0.000
	lavaan	—	—	5.284
28%	Integrated	508.00	−285.73	1.834
	NonIntegrated	503.68	−257.19	1.787
37%	Integrated	499.62	−246.98	2.567
	NonIntegrated	495.85	−236.26	2.551
50%	Integrated	486.30	−240.78	4.088
	NonIntegrated	483.29	−209.78	4.031

average MSE 2.36 is far larger than that of integrated model. Moreover, for all the censoring cases and non-censoring cases, the MSE due to the integrated model remains smaller than the nonIntegrated model.

3.2 NonIntegrated model as the data-generating model

This section is devoted to the simulation study when the data are generated from the nonIntegrated model (8). We use this model to generate 100 simulated datasets in which the data generation scheme was very similar to what have been discussed in Section 3.1. After generating the dataset according to a nonIntegrated model, we fit both integrated model and nonIntegrated model. We provide the summary of the results in Table 1. We note that, for instance, when there are about 28% of the data are right censored and a nonIntegrated model is fitted in the generated datasets, the average DIC is 503.68 and when the integrated model is fitted then the average DIC is 508.00. Hence, it can be concluded that even though the integrated model does not provide a better fit the difference is, however, very small to distinguish unlike the case when the data-generating model is the integrated model. This phenomenon is evident in the other results of DIC, LPML and MSE in Table 1.

3.3 Sensitivity analysis

The purpose of this example to examine the effect in the performance of our proposed integrated model under different fixed values of the variance parameters σ_{u_1} and σ_{u_2}. We generate the data in the same way as in Section 3.1. The censoring distribution parameters are set in such a way that the average censoring for 100 simulated data is about 25%. The other priors were similar to what we had in the previous section. Table 2 presents the results under different set of values of σ_{u_1} and σ_{u_2}. One can notice that even though we vary the fix values of σ_{u_1} and σ_{u_2}, we see a little deviation of the results in terms of the DIC, LPML and MSE values of the fitted integrated model. This follows that when the values of σ_{u_1} and σ_{u_2} are within the range of (0, 2), then the integrated model is not affected by the fixed values of these parameters.

Table 2. DIC, LPML and MSE of the integrated model for simulated data under different values of $\sigma_{u_1}^2$ and $\sigma_{u_2}^2$, censoring rate =24%

$(\sigma_{u1}^2, \sigma_{u2}^2)$	DIC	LPML	MSE
(0.25, 0.25)	305.64	−290.84	0.0419
(0.50, 0.50)	305.73	−290.33	0.0417
(0.75, 0.75)	306.20	−289.04	0.0422
(1.00, 1.00)	306.33	−291.35	0.0417
(1.50, 1.50)	307.32	−289.34	0.0428
(2.00, 2.00)	307.34	−289.43	0.0423

4 Circadian genes from TCGA

In TCGA data, among available omics expressions, the DNA copy number changes are collected via SNP-arrays and array comparative genomic hybridization and for breast cancer data, only the first kind is available via the R package *TCGA2STAT* (Wan *et al.*, 2015). TCGA provides the gene expression data in several different forms and among them we have considered the one, which is measured via RNA-sequencing technology preprocessed using the first pipeline and normalized to get continuous measurements, which is known as Reads Per Kilobase Million. The original data are the version-

Table 3. Circadian genes used for TCGA data analysis

Genes	Description
CRY1	Belongs to the flavoproteins superfamily that exists in all kingdoms of life and act as light-independent inhibitors of CLOCK-BMAL1 components of the circadian clock
CRY2	Belong to the flavoproteins superfamily that exists in all kingdoms of life and act as light-independent inhibitors of CLOCK-BMAL1 components of the circadian clock
CSNK1E	The protein encoded by this gene is a serine/threonine protein kinase and a member of the casein kinase I protein family, whose members have been implicated in the control of cytoplasmic and nuclear processes, including DNA replication and repair
DEC1	Transcriptional repressor involved in the regulation of the circadian rhythm by negatively regulating the activity of the clock genes and clock-controlled genes
MT2	Is a member of the metallothionein family of genes. Proteins encoded by this gene family are low in molecular weight, are cysteine-rich, lack aromatic residues and bind divalent heavy metal ions, altering the intracellular concentration of heavy metals in the cell
NPAS2	A protein-coding gene and a transcriptional activator, which forms a core component of the circadian clock
PER1	Encodes the period circadian protein homolog 1 protein in humans
PER2	A member of the Period family of genes and is expressed in a circadian pattern in the suprachiasmatic nucleus (SCN)
PER3	Expressed in a circadian pattern in the SCN, the primary circadian pacemaker in the mammalian brain
TIMELESS	Is notable for its role in Drosophila for encoding TIM, an essential protein that regulates circadian rhythm

stamped standardized datasets hosted and maintained by the Broad Institute GDAC Firehose.

Our study focuses on the circadian genes and their effects on the patients' survival. We collected 10 such gene expressions (Table 3) with the corresponding observed survival components, the age and the gender of 68 glioblastoma tumor samples and 364 breast tumor samples.

4.1 Glioblastoma cancer data analysis

Glioblastoma, also known as glioblastoma multiforme or grade IV astrocytoma, is a fast-growing, aggressive type of central nervous system tumor that forms on the supportive tissue of the brain and it is the most common grade IV brain cancer. In 2018, more than 23 000 Americans were estimated to have been diagnosed and among them 16 000 were estimated to have died from brain and other nervous system cancers (Siegel *et al.*, 2018). Glioblastoma accounts for about 15% of all brain tumors and occurs in adults between the ages of 45–70 years. Among the available data about 27% are right censored.

In addition, in this analysis, we consider the gender and the age of the patients as the external predictors on the survival time. As an exploratory analysis, we fit a log-normal AFT regression of the survival times of the individuals on their age. Figure 3 displays the residuals and the Q–Q plot of those residuals. The residual plot shows that there is no clear pattern in the residuals. Furthermore, the Q–Q plot establishes that the log-normal assumption on the residual distribution is adequate.

We specify the following values for the prior distribution parameters for various parameters. For example, we set $\beta_{t0} = 0$, Σ_β is the unit variance-covariance matrix, i.e. the diagonals are set to 1 and the off-diagonals are 0. We set, $\alpha_{t0} = 0$, $\sigma^2_{\alpha_t} = 1$. Similarly, the

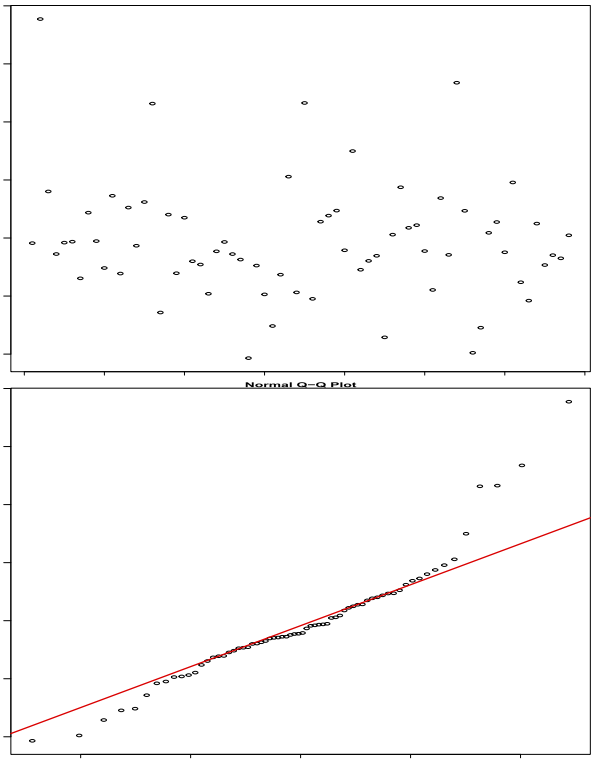

Fig. 3. Top panel: residual plot of the log-normal regression of the survival time on the age of the individuals. Bottom panel: Q–Q plot of those residual against the normal distribution

standard normal distribution is placed as the prior distributions of $\alpha_{u_{1k}}, \alpha_{u_{2l}}, \phi_t, \phi_{u_{1k}}$, and $\phi_{u_{2l}}$, $k = 1, \ldots, q_1$, $l = 1, \ldots, q_2$. For our experiment $q_1 = q_2 = 10$.

We provide the goodness of fit results in Table 4 and we note that the results of DIC and LPML suggest the superior performance of the proposed integrated approach compared to the traditional one. For example, when the SEM is fitted to the data the DIC is −244.93, which is lower than the DIC, 213.20, when the nonIntegrated model is fitted. Furthermore, in Figure 4, for two randomly selected individuals, we depict the survival probabilities computed using the two methods on the Kaplan–Meier plot.

4.1.1 Comparison with the existing method

In this section, we provide a brief study to find the performance of our proposed integrated structural equation model in comparison to the existing iBAG method (Wang *et al.*, 2013). The experiment is carried out on the Glioblastoma dataset. We note that, iBAG method is primarily developed to assess the individual gene effect on the clinical outcome while considering the underlying relationship between the different high-dimensional omics data platforms, such as methylation and mRNA expressions. To this end, this method employs a high-dimensional Bayesian variable selection in the fitting of the model. In contrast, in this article, the proposed structural equation method is examined only for circadian genes, which are responsible for exhibiting time dependent behavior across 24 h of each day, i.e. we are interested in explaining the relationship between a particular trait and the survival of the cancer patients. Since feature selection is not the primary interest of our study, a direct comparison is beyond the scope of this article.

Nevertheless, in this example, we present a comparative study of both the methods. The computation for the iBAG method is carried out using the code given in Wang *et al.* (2013). When fitting iBAG to the Glioblastoma dataset considered here, we replace the methylation expressions and the mRNA expressions with the CNV and the RNQSeq data, respectively. The available program does not

Table 4. Goodness of fit for the integrated and nonIntegrated models in glioblastoma data

Method	DIC	LPML	MSE
Integrated	−244.93	−175.58	0.230
NonIntegrated	213.20	−303.44	0.459
iBAG	—	—	0.392

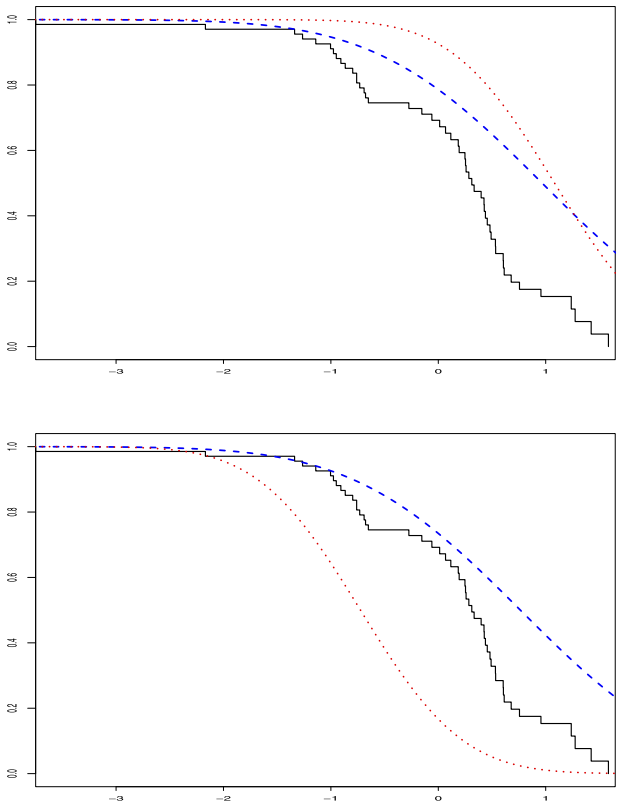

Fig. 4. Survival functions for randomly selected two individuals for glioblastoma cancer dataset. Solid (black): the Kaplan–Meier plot, dashed (blue): integrated structural equation model and dotted (red): nonIntegrated model

provide DIC and LPML for the fitted model. Hence, the MSE is computed on the uncensored time points and is given in Table 4. One can note that, the MSE due to our proposed integrated method is 0.230 and the same for iBAG method is 0.392. This concludes that the propose SEM method remains superior in terms of the prediction performance.

4.2 Breast cancer data analysis

Breast cancer is one of the most common cancers with a massive number of cases reported. For instance, in 2018, more than 268 000 Americans were estimated to have been diagnosed and 41 000 were estimated to have died from breast cancer related tumors (Siegel *et al.*, 2018). This heterogeneous disease is categorized into three groups, such as the oestrogen receptor group, the HER2 amplified group and the triple-negative breast cancers or the basal-like breast cancers (Network, 2012). Among them, we consider the information of 364 breast tumor samples with their survival data from TCGA. We observe that at least 82% data are right censored. In the analysis, we consider the age variable as a covariate effect on the survival time.

We present the goodness of fit results in Table 5 and we notice that the DIC due to our proposed method is 1077.38, which is less than the DIC 1104.87 due to the nonIntegrated model. This indicates that the proposed integrated model provides a better fit to the

Table 5. Goodness of fit for the integrated and nonIntegrated models in breast cancer data

Method	DIC	LPML
Integrated	1077.38	−277.24
NonIntegrated	1104.87	−384.10

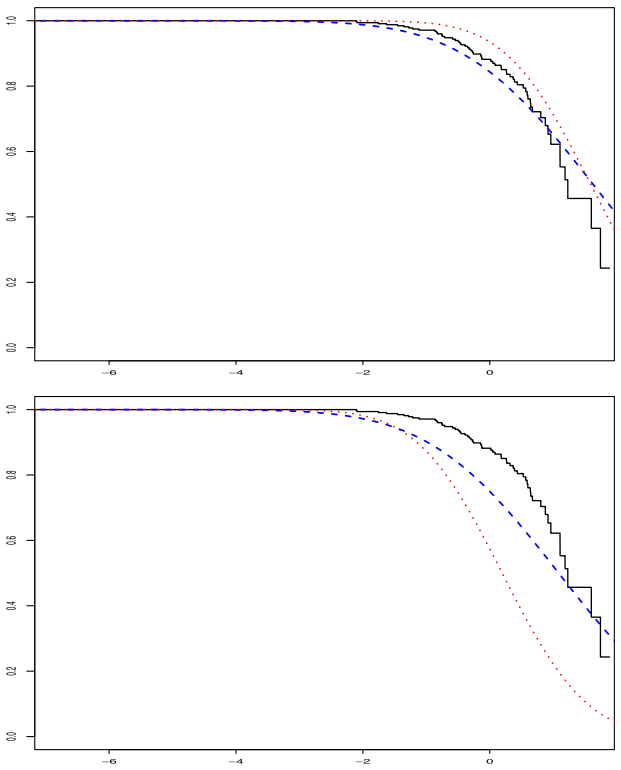

Fig. 5. Survival functions for randomly selected two individuals for breast cancer dataset. Solid (black): the Kaplan–Meier plot, dashed (blue): integrated structural equation model and dotted (red): nonIntegrated model

breast cancer data. This is also confirmed by the LPML numbers obtained by fitting the different models to data. In Figure 5, for two randomly selected individuals, we depict the survival probabilities computed using the two methods on the Kaplan–Meier plot.

5 Conclusion

In this article, we have proposed a simple Bayesian SEM technique to integrate the information from different omics platform. We have shown that the proposed SEM technique provides improved survival prediction and better fits to the data compared to the traditional approach. Our focus in this article is concentrated on circadian genes only. Toward this end, the sole intention of the proposed method is to capture the biological system in order to predict the patient survival when the circadian genes are of the interest.

Nonetheless, when a large number of gene expressions is under consideration and we have only limited number of patient samples then a sophisticated variable selection method needs to be implemented which will also have the ability to detect the effect of a single gene on the clinical outcome.

In a very general setup, we can allow a latent variable for each gene and use appropriate priors to borrow strength. This will be an over parameterized model with huge number of random effects and due to their correlations the computation will be extremely slow and expensive. The remedy is to categorize (cluster or group) the genes according to their functions and use a latent variable corresponding

to each of these categories. In our applications, we are working only with circadian genes, which can be treated as a single category and hence we have specified a single latent variable corresponding to it. Extension to multi-category models will be done in future research using clustered models.

We have specified a non-informative prior on σ_t^2. It is worth to mention that an Inverse Gamma prior would also maintain the conjugacy. However, our study shows that, imposing a suitable prior on all other variance parameters results in similar superior performance of the proposed structural equation based integrated modeling, which is evident from the analysis given in the Supplementary Material. Hence, choice of appropriate priors for those parameters is kept for future studies.

The two platforms, we have considered here are RNAseq and CNV. In these regressions, we separately regress the corresponding expressions on two separate latent variables for each gene. Hence, we have assumed that those regressions are conditionally independent from each other's. If a particular application violates this assumption caution should be exercised.

We assume our model specification to be fully parametric. As a starting approach, the log-normal model is assumed here. A Weibull model or a Gamma model is also possible to fit. However, all of these distributions have similar tail property. Moreover, we examine the residual plots of the log-normal models (included in Section 4.1 for the age variable and in the Supplementary Material for few genes), which are satisfactory for a log-normal assumption. Nevertheless, one possible extension, as indicated by Wong *et al.* (2018), is to consider non-parametric models, which is due for the future research. The theoretical properties are also of future interests.

The latent variables, which are key components of the proposed model are platform specific, i.e. each platform expression is regulated by a single latent variable, which is sufficient for circadian oscillation characteristics. Using this and using the log-normal AFT model, we have developed the structural equation model to predict the clinical outcome survival. The log-normal AFT model has been shown adequately fitted to the TCGA data considered here. In our examples, we showed that the proposed model outperformed independent models. However, one must be aware that if any or some of the assumptions are not satisfied then the model should be tuned accordingly.

Acknowledgements

We are grateful to the editor, the associate editor Inanc Birol and the three anonymous referees whose valuable comments have considerably improved this article.

Funding

The research reported in this article was supported by grants from the National Cancer Institute [R01-CA194391]; and National Science Foundation [CCF-1934904].

Conflict of Interest: none declared.

References

Andreani,T.S. *et al.* (2015) Genetics of circadian rhythms. *Sleep Med. Clin.*, **10**, 413–421.

Bentler,P.M. and Weeks,D.G. (1980) Linear structural equations with latent variables. *Psychometrika*, **45**, 289–308.

Bollen,K.A. and Davis,W.R. (2009) Two rules of identification for structural equation models. *Struct. Equ. Modeling*, **16**, 523–536.

Bonato,V. *et al.* (2011) Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, **27**, 359–367.

Brown,E.R. *et al.* (2005) A flexible B-spline model for multiple longitudinal biomarkers and survival. *Biometrics*, **61**, 64–73.

Chu,S.H. and Huang,Y.-T. (2017) Integrated genomic analysis of biological gene sets with applications in lung cancer prognosis. *BMC Bioinformatics*, **18**, 336.

Daemen,A. *et al.* (2009) A kernel-based integration of genome-wide data for clinical decision support. *Genome Med.*, **1**, 39.

Davis,S. and Mirick,D.K. (2006) Circadian disruption, shift work and the risk of cancer: a summary of the evidence and studies in Seattle. *Cancer Causes Control*, **17**, 539–545.

Fu,L. and Kettner,N.M. (2013) The circadian clock in cancer development and therapy. *Prog. Mol. Biol. Transl. Sci.*, **119**, 221–282.

Geisser,S. and Eddy,W.F. (1979) A predictive approach to model selection. *J. Am. Stat. Assoc.*, **74**, 153–160.

Gelfand,A.E. *et al.* (1992) Model determination using predictive distributions with implementation via sampling-based methods. *Technical report*. Department of Statistics, Stanford University.

Gomez-Cabrero,D. *et al.* (2014) Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* **8**, I1.

Hamid,J.S. *et al.* (2009) Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics*, **2009**, 869093.

Heckman,J.J. and Vytlacil,E. (2005) Structural equations, treatment effects, and econometric policy evaluation 1. *Econometrica*, **73**, 669–738.

Huang,S. *et al.* (2017) More is better: recent progress in multi-omics data integration methods. *Front. Genet.*, **8**, 84.

Ibrahim,J.G. and Laud,P.W. (1994) A predictive approach to the analysis of designed experiments. *J. Am. Stat. Assoc.*, **89**, 309–319.

Ibrahim,J.G. *et al.* (2005) *Bayesian Survival Analysis. Wiley Online Library*. New York, New York, USA.

Larsen,K. (2005) The Cox proportional hazards model with a continuous latent variable measured by multiple binary indicators. *Biometrics*, **61**, 1049–1055.

Naliboff,B.D. *et al.* (2012) Gastrointestinal and psychological mediators of health-related quality of life in IBS and IBD: a structural equation modeling analysis. *Am. J. Gastroenterol.*, **107**, 451.

Network,C.G.A. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.

Palomo,J. *et al.* (2007) Bayesian structural equation modeling. In: Sik-Yum,L. ed. *Handbook of Latent Variable and Related Models*. North Holland, Amsterdam, The Netherlands, pp. 163–188.

Quintana,S.M. and Maxwell,S.E. (1999) Implications of recent developments in structural equation modeling for counseling psychology. *Couns. Psychol.*, **27**, 485–527.

Rizopoulos,D. and Ghosh,P. (2011) A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat. Med.*, **30**, 1366–1380.

Rosseel,Y. (2012) lavaan: an R package for structural equation modeling. *J. Stat. Softw.*, **48**, 1–36.

Sahar,S. and Sassone-Corsi,P. (2009) Metabolism and cancer: the circadian clock connection. *Nat.Rev. Cancer*, **9**, 886–896.

Shen,R. *et al.* (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.

Siegel,R.L. *et al.* (2018) Cancer statistics, 2018. *CA Cancer J. Clin.*, **68**, 7–30.

Song,X.-Y. and Lee,S.-Y. (2012) A tutorial on the Bayesian approach for analyzing structural equation models. *J. Math. Psychol.*, **56**, 135–148.

Spiegelhalter,D.J. *et al.* (2002) Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Series B*, **64**, 583–639.

Stoolmiller,M. and Snyder,J. (2006) Modeling heterogeneity in social interaction processes using multilevel survival analysis. *Psychol. Methods*, **11**, 164–177.

Tanner,M.A. and Wong,W.H. (1987) The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.*, **82**, 528–540.

Tseng,G. *et al.* (2015) *Integrating Omics Data*. Cambridge University Press, Cambridge.

Vaske,C.J. *et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.

Wan,Y.-W. *et al.* (2015) *TCGA2STAT: Simple TCGA Data Access for Integrated Statistical Analysis in R*. R package version 1.2. https://doi.org/10.1186/1752-0509-8-S2-I1.

Wang,W. *et al.* (2013) iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics*, **29**, 149–159.

Weinstein,J.N. *et al.*; The Cancer Genome Atlas Research Network. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

Wong,K.Y. *et al.* (2018) Efficient estimation for semiparametric structural equation models with censored data. *J. Am. Stat. Assoc.*, **113**, 893–905.