BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

Combining heterogeneous subgroups with graph-structured variable selection priors for Cox regression

Katrin Madjar[1*] , Manuela Zucknick[2], Katja Ickstadt[1] and Jörg Rahnenführer[1]

*Correspondence:
madjar@statistik.tu-dortmund.de
[1] Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany
Full list of author information is available at the end of the article

Abstract

Background: Important objectives in cancer research are the prediction of a patient's risk based on molecular measurements such as gene expression data and the identification of new prognostic biomarkers (e.g. genes). In clinical practice, this is often challenging because patient cohorts are typically small and can be heterogeneous. In classical subgroup analysis, a separate prediction model is fitted using only the data of one specific cohort. However, this can lead to a loss of power when the sample size is small. Simple pooling of all cohorts, on the other hand, can lead to biased results, especially when the cohorts are heterogeneous.

Results: We propose a new Bayesian approach suitable for continuous molecular measurements and survival outcome that identifies the important predictors and provides a separate risk prediction model for each cohort. It allows sharing information between cohorts to increase power by assuming a graph linking predictors within and across different cohorts. The graph helps to identify pathways of functionally related genes and genes that are simultaneously prognostic in different cohorts.

Conclusions: Results demonstrate that our proposed approach is superior to the standard approaches in terms of prediction performance and increased power in variable selection when the sample size is small.

Keywords: Bayesian variable selection, Cox proportional hazards model, Gaussian graphical model, Markov random field prior, Heterogeneous cohorts, Subgroup analysis

Background

In clinical research, molecular measurements such as gene expression data play an important role in the diagnosis and prediction of a disease outcome, such as time-to-event endpoint. In general, the number of molecular predictors is larger than the sample size ("$p > n$ problem") and typically only a small number of genes is associated with the outcome while the rest is noise. Thus, important objectives in statistical modeling are good prediction performance and variable selection to obtain a subset of prognostic predictors.

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 2 of 29

In the Bayesian framework, different types of variable selection priors have been proposed also with application to the Bayesian Cox model. One common choice is the use of shrinkage priors such as the Bayesian lasso as an analog to the frequentist penalized likelihood approach [20, 26, 41]. A popular alternative are "spike-and-slab" priors that use latent indicators for variable selection and a mixture distribution for the regression coefficients [14, 35]. In general, the regression coefficients are modeled independently. However, with applications to molecular data, it can be reasonable to consider structural information between covariates, since the effect on a clinical outcome is typically not caused by single genes acting in isolation, but rather by changes in a regulatory or functional pathway of interacting genes. Several authors have dealt with this problem by using a Markov random field (MRF) prior to incorporate structural information on the relationships among the covariates into variable selection [21, 28, 33, 34]. Alternatively, Chakraborty and Lozano [5] propose a Graph Laplacian prior for modeling the dependence structure between the regression coefficients through their precision matrix.

When the data are heterogeneous and consists of known subpopulations with possibly different dependence structures, estimating one joint graphical model would hide the underlying heterogeneity while estimating separate models for each subpopulation would neglect common structure. For this situation, Danaher et al. [8] use an extension of the frequentist graphical lasso with either a group or fused lasso type penalty for joint structure learning. Saegusa and Shojaie [30] propose a weighted Laplacian shrinkage penalty where the weights represent the degree of similarity between subpopulations. Bayesian approaches for sharing common structure in the joint inference of multiple graphical models have also been developed [24, 27, 40]. Peterson et al. [27] use an MRF prior for the graph structures with pairwise similarities between different graphs. However, all these methods have in common that they focus on structure learning only and do not take into account the relationship between (structured) covariates and a clinical outcome as in the context of regression modeling.

We consider the situation that molecular measurements and a survival outcome are available for various, possibly heterogeneous patient subgroups or cohorts such as in a multicenter study. In the following, we use the term "subgroup" for different pre-known patient cohorts or data sets. In classical subgroup analysis, only the data of the subgroup of interest is used to build a risk prediction model for this specific subgroup. This may lead to a loss of power or unstable results with high variance especially in small subgroups. Thus, it is tempting to simply pool all data to increase the sample size. This approach, however, can result in biased estimates when the subgroups are heterogeneous regarding their effects and subgroup-specific effects may get lost. We aim at sharing information between subgroups to increase power when this is supported by the data. Our approach provides a separate risk prediction model for each subgroup that allows the identification of common as well as subgroup-specific effects and has improved prediction accuracy and variable selection power compared to the two standard approaches.

Some frequentist approaches tackle this problem by suggesting a penalized Cox regression model with a weighted version of the partial likelihood that includes patients of all subgroups but assigns them (individual) weights. Weyer and Binder [37] propose the use of fixed weights. This idea is extended by Richter et al. [29] using model-based optimization for tuning of the weights to obtain the best combination of fixed weights regarding

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 3 of 29

prediction accuracy. [23] estimate individual weights from the data such that they represent the probability of belonging to a specific subgroup.

In this paper, we use a Bayesian approach and borrow information across subgroups through graph-structured selection priors instead of weights in the likelihood. We propose an extension of the Bayesian Cox model with "spike-and-slab" prior for variable selection by Treppmann et al. [35] in the sense that we incorporate graph information between covariates into variable selection via an MRF prior instead of modeling the regression coefficients independently. The graph is not known a priori and inferred simultaneously with the important predictors. Its structure can be partitioned into subgraphs linking covariates within or across different subgroups. Thus, representing conditional dependencies between genes (i.e. pathways) and similarities between subgroups by genes being simultaneously prognostic in different subgroups.

Methods

First, the general methods are described that are required for our proposed Bayesian model introduced later in this section.

The Bayesian Cox proportional hazards model

Assume the observed data of patient m consist of the tuple $(\tilde{t}_m, \delta_m)$ and the covariate vector $\boldsymbol{x}_m = (x_{m1}, \ldots, x_{mp})' \in \mathbb{R}^p$, $m = 1, \ldots, n$. $\boldsymbol{x} \in \mathbb{R}^{n \times p}$ is the matrix of (genomic) covariates. $\tilde{t}_m = \min(T_m, C_m)$ denotes the observed time of patient m, with T_m the event time and C_m the censoring time. $\delta_m = \mathbb{1}(T_m \leq C_m)$ indicates whether a patient experienced an event ($\delta_m = 1$) or was right-censored ($\delta_m = 0$).

The Cox proportional hazards model [7] models the hazard rate $h(t|\boldsymbol{x}_m)$ of an individual m at time t. It consists of two terms, the non-parametric baseline hazard rate $h_0(t)$ and a parametric form of the covariate effect:

$$h(t|\boldsymbol{x}_m) = h_0(t) \cdot \exp(\boldsymbol{\beta}'\boldsymbol{x}_m) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i x_{mi}\right),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the unknown parameter vector that represents the strength of influence of the covariates on the hazard rate.

Under the Cox model, the joint survival probability of n patients given $\boldsymbol{x}$ is

$$P(\tilde{\boldsymbol{T}} > \tilde{\boldsymbol{t}}|\boldsymbol{x}, \boldsymbol{\beta}, H_0) = \exp\left(-\sum_{m=1}^{n} \exp(\boldsymbol{\beta}'\boldsymbol{x}_m)H_0(\tilde{t}_m)\right),$$

where $\tilde{\boldsymbol{t}}$ is the vector of the individual observed times for all patients and $\tilde{\boldsymbol{T}}$ the vector of corresponding random variables. One of the most popular choices for the cumulative baseline hazard function $H_0(t)$ is a gamma process prior

$$H_0 \sim \mathcal{GP}(a_0 H^*, a_0),$$

where $H^*(t)$ is an increasing function with $H^*(0) = 0$. H^* can be considered as an initial guess of H_0 and $a_0 > 0$ describes the weight that is given to $H^*(t)$ [20]. Lee et al. [20] propose a Weibull distribution $H^*(t) = \eta t^\kappa$ with fixed hyperparameters η and κ. Following Zucknick et al. [41], we obtain estimates of η and κ from the training data by fitting a

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 4 of 29

parametric Weibull model without covariates to the survival data. We choose $a_0 = 2$ in accordance with the authors.

In practice the presence of ties is very common, leading to the grouped data likelihood described in Ibrahim et al. [17, chapter 3.2.2]. A finite partition of the time axis is constructed with $0 = c_0 < c_1 < \ldots < c_J$ and $c_J > \tilde{t}_m$ for all $m = 1, \ldots, n$. The observed time $\tilde{t}_m$ of patient m falls in one of the J disjoint intervals $I_g = (c_{g-1}, c_g]$, $g = 1, \ldots, J$. Assume the observed data $\mathfrak{D} = \{(\boldsymbol{x}, \mathcal{R}_g, \mathcal{D}_g) : g = 1, \ldots, J\}$ are grouped within I_g, where $\mathcal{R}_g$ and $\mathcal{D}_g$ are the risk and failure sets corresponding to interval g. Let $h_g = H_0(c_g) - H_0(c_{g-1})$ be the increment in the cumulative baseline hazard in interval I_g, $g = 1, \ldots, J$. From the gamma process prior of H_0 follows that the h_g's have independent gamma distributions

$$h_g \sim \mathcal{G}(\alpha_{0,g} - \alpha_{0,g-1}, a_0), \quad \text{with} \quad \alpha_{0,g} = a_0 H^*(c_g).$$

The conditional probability that the observed time of patient m falls in interval I_g is given by

$$P(\tilde{T}_m \in I_g | \boldsymbol{h}) = \exp\left(-\exp(\boldsymbol{\beta}'\boldsymbol{x}_m) \sum_{j=1}^{g-1} h_j\right) \cdot \left[1 - \exp\left(-h_g \exp(\boldsymbol{\beta}'\boldsymbol{x}_m)\right)\right],$$

with $\boldsymbol{h} = (h_1, \ldots, h_J)'$. The resulting grouped data likelihood is defined as

$$\begin{aligned} &L(\mathfrak{D}|\boldsymbol{\beta}, \boldsymbol{h}) \\ &\propto \prod_{g=1}^{J} \left[\exp\left(-h_g \sum_{k \in \mathcal{R}_g - \mathcal{D}_g} \exp(\boldsymbol{\beta}'\boldsymbol{x}_k)\right) \prod_{l \in \mathcal{D}_g} \left[1 - \exp\left(-h_g \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)\right)\right]\right] \end{aligned}$$

[17, chapter 3.2.2].

Stochastic search variable selection

The stochastic search variable selection (SSVS) procedure by George and McCulloch [14] uses latent indicators for variable selection and models the regression coefficients as a mixture of two normal distributions with different variances

$$\beta_i | \gamma_i \sim (1 - \gamma_i) \cdot \mathcal{N}(0, \tau_i^2) + \gamma_i \cdot \mathcal{N}(0, c_i^2 \tau_i^2), \quad i = 1, \ldots, p.$$

This prior allows the β_i's to shrink towards zero. Due to the shape of the two-component mixture distribution, it is called *spike-and-slab prior*. The latent variable γ_i indicates the inclusion ($\gamma_i = 1$) or exclusion ($\gamma_i = 0$) of the i-th variable and specifies the variance of the normal distribution. τ_i (> 0) is set small so that β_i is likely to be close to zero if $\gamma_i = 0$. c_i (> 1) is chosen sufficiently large to inflate the coefficients of selected variables and to make their posterior mean values likely to be non-zero. In general, the variances of the regression coefficients are assumed to be constant: $\tau_i \equiv \tau$ and $c_i \equiv c$ for all $i = 1, \ldots, p$.

The standard prior for $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ is a product of independent Bernoulli distributions

$$p(\boldsymbol{\gamma}) = \prod_{i=1}^{p} \pi_\gamma^{\gamma_i} \cdot (1 - \pi_\gamma)^{1 - \gamma_i},$$

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 5 of 29

with prior inclusion probability $\pi_\gamma = P(\gamma_i = 1)$. Typically, these prior inclusion probabilities are chosen to be the same for all variables and often with π_γ set to a fixed value.

Graphical models

A graphical model is a statistical model that is associated with a graph summarizing the dependence structure in the data. The nodes of a graph represent the random variables of interest and the edges of a graph describe conditional dependencies among the variables. Structure learning implies the estimation of an unknown graph. Recent applications are mainly driven by biological problems that involve the reconstruction of gene regulatory networks and the identification of pathways of functionally related genes from their expression levels. A graph is called *undirected,* when its edges are unordered pairs of nodes instead of ordered pairs with edges pointing from one node to the other (*directed* graph). When the variables are continuous measurements and assumed to be multivariate normal a common choice are Gaussian models [11].

We assume that the vector of random variables $X_m = (X_{m1}, \ldots, X_{mp})'$ for patient m, $m = 1, \ldots, n$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ. The inverse of the covariance matrix is referred to as precision matrix $\Sigma^{-1} = \Omega = (\omega_{ij})_{i,j=1,\ldots,p}$, with Ω symmetric and positive definite. Let $X \in \mathbb{R}^{n \times p}$ be the data matrix consisting of n independent patients and $S = \frac{1}{n} X'X$ the sample covariance matrix.

In graphical models, a graph $\widetilde{G}$ is used to represent conditional dependence relationships among random variables X. Let $\widetilde{G} = (V, E)$ be an undirected graph, where $V = \{1, \ldots, p\}$ is a set of nodes (e.g. genes) and $E \subset V \times V$ is a set of edges (e.g. relations between genes) with edge $(i, j) \in E \Leftrightarrow (j, i) \in E$. $\widetilde{G}$ can be indexed by a set of $p(p-1)/2$ binary variables $G = (g_{ij})_{i<j} \in \{0, 1\}^{p \times p}$ with $g_{ij} = 1$ or 0 when edge (i, j) belongs to E or not. The symmetric matrix G is termed adjacency matrix representation of the graph. The graph structure implies constraints on the precision matrix Ω such that $g_{ij} = 0 \Leftrightarrow (i, j) \notin E \Leftrightarrow \omega_{ij} = 0$, meaning that variables i and j are conditionally independent given all remaining variables [11, 36].

We use the approach for structure learning by Wang [36] that is based on continuous spike-and-slab priors for the elements of the precision matrix and latent indicators for the graph structure. The approach induces sparsity and is efficient due to a block Gibbs sampler and no approximation of the normalizing constant. The corresponding hierarchical model is defined as

$$p(\Omega|G, \theta) = C(G, v_0, v_1, \lambda)^{-1} \prod_{i<j} \mathcal{N}(\omega_{ij}|0, v_{g_{ij}}^2) \prod_i \text{Exp}(\omega_{ii}|\frac{\lambda}{2}) \mathbb{1}_{\{\Omega \in \mathcal{M}^+\}}$$

$$p(G|\theta) = C(\theta)^{-1} C(G, v_0, v_1, \lambda) \prod_{i<j} \left(\pi_G^{g_{ij}} (1 - \pi_G)^{1-g_{ij}} \right),$$

where $C(G, v_0, v_1, \lambda)$ and $C(\theta)$ are the normalizing constants, and $\theta = \{v_0, v_1, \lambda, \pi_G\}$ is the set of all parameters with $v_0 > 0$ small, $v_1 > 0$ large, $\lambda > 0$ and $\pi_G \in (0, 1)$. $\mathbb{1}_{\{\Omega \in \mathcal{M}^+\}}$ restricts the prior to the space of symmetric-positive definite matrices. A small value for v_0 ($g_{ij} = 0$) means that ω_{ij} is small enough to bet set to zero. A large value for v_1 ($g_{ij} = 1$) allows ω_{ij} to be substantially different from zero. The binary latent variables $G = (g_{ij})_{i<j} \in \{0, 1\}^{p(p-1)/2}$ serve as edge inclusion indicators. Wang [36] proposes the

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 6 of 29

following fixed hyperparameters $\pi_G = \frac{2}{p-1}$, $\nu_0 \geq 0.01$, $\nu_1 \leq 10$ and $\lambda = 1$ as resulting in good convergence and mixing.

The proposed Bayesian subgroup model

We assume the entire data consists of S predefined subgroups of patients (different cohorts or data sets), where for each patient the subgroup affiliation is known and unique. This information, which specific subgroup a patient belongs to, is available in the data.

Likelihood

Let $X_s \in \mathbb{R}^{n_s \times p}$ be the gene expression (covariate) matrix for subgroup s, $s = 1, \ldots, S$, consisting of n_s independent and identically distributed observations. For patient m in subgroup s the vector of random variables $X_{s,m} = (X_{s,m1}, \ldots, X_{s,mp})'$ is assumed to follow a multivariate normal distribution with mean vector $\mathbf{0}$ and unknown precision matrix $\mathbf{\Omega}_{ss} = \mathbf{\Sigma}_s^{-1}$, $m = 1, \ldots, n_s$.

We consider the outcome $Y_s = (Y_{s,1}, \ldots, Y_{s,n_s})'$ with $Y_{s,m} = (\tilde{T}_{s,m}, \delta_{s,m})$ as well as the predictors X_s, to be random variables. Thus, the likelihood for subgroup s is the joint distribution $p(Y_s, X_s) = p(Y_s | X_s) \cdot p(X_s)$. The conditional distribution $p(Y_s | X_s)$ corresponds to the grouped data likelihood of the Bayesian Cox proportional hazards model at the beginning of this section [20] for subgroup s

$$L(\mathfrak{D}_s | \boldsymbol{\beta}_s, \boldsymbol{h}_s)$$

$$\propto \prod_{g=1}^{J_s} \left[\exp\left(-h_{s,g} \sum_{k \in \mathcal{R}_{s,g} - \mathcal{D}_{s,g}} \exp(\boldsymbol{\beta}_s' \boldsymbol{x}_{s,k}) \right) \prod_{l \in \mathcal{D}_{s,g}} \left[1 - \exp\left(-h_{s,g} \exp(\boldsymbol{\beta}_s' \boldsymbol{x}_{s,l}) \right) \right] \right],$$

where $\mathfrak{D}_s = \{(\boldsymbol{x}_s, \mathcal{R}_{s,g}, \mathcal{D}_{s,g}) : g = 1, \ldots, J_s\}$ are the observed data in subgroup s, with $\mathcal{R}_{s,g}$ the risk and $\mathcal{D}_{s,g}$ the failure sets corresponding to interval $I_{s,g} = (c_{s,g-1}, c_{s,g}]$, $g = 1, \ldots, J_s$. The increment in the cumulative baseline hazard for subgroup s in interval $I_{s,g}$ is termed $h_{s,g} = H_0(c_{s,g}) - H_0(c_{s,g-1})$. $\boldsymbol{\beta}_s$ is the p-dimensional vector of regression coefficients for subgroup s.

The marginal distribution of X_s is multivariate normal with $S_s = X_s' X_s$

$$p(X_s | \mathbf{\Omega}_{ss}) \propto \prod_{m=1}^{n_s} |\mathbf{\Omega}_{ss}|^{1/2} \exp\left(-\frac{1}{2} X_{s,m}' \mathbf{\Omega}_{ss} X_{s,m} \right)$$

$$= |\mathbf{\Omega}_{ss}|^{n_s/2} \exp\left(-\frac{1}{2} \underbrace{\sum_{m=1}^{n_s} X_{s,m}' \mathbf{\Omega}_{ss} X_{s,m}}_{=\operatorname{tr}(S_s \mathbf{\Omega}_{ss})} \right).$$

The joint likelihood across all subgroups is the product of the subgroup likelihoods

$$\prod_{s=1}^{S} L(\mathfrak{D}_s | \boldsymbol{\beta}_s, \boldsymbol{h}_s) \cdot p(X_s | \mathbf{\Omega}_{ss}).$$

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 7 of 29

Prior specifications

Prior on the parameters h_s and β_s of the Cox model

The prior for the increment in the cumulative baseline hazard in subgroup s follows independent gamma distributions

$$h_{s,g} \sim \mathcal{G}(a_0(H^*(c_{s,g}) - H^*(c_{s,g-1})), a_0),$$

with a Weibull distribution $H^*(c_{s,g}) = \eta_s c_{s,g}^{\kappa_s}$, $g = 1, \ldots, J_s$, $s = 1, \ldots, S$ [20]. We choose the hyperparameters a_0, η_s and κ_s to be fixed and in accordance with Lee et al. [20] and Zucknick et al. [41]. We set $a_0 = 2$ and estimate the hyperparameters η_s and κ_s from the (training) data by fitting a parametric Weibull model without covariates to the survival data of subgroup s.

We perform variable selection using the SSVS approach by George and McCulloch [14] described earlier in this section. The prior of the regression coefficients $\beta_{s,i}$ in subgroup s conditional on the latent indicator $\gamma_{s,i}$ is defined as a mixture of two normal distributions with small (τ^2) and large ($c^2\tau^2$) variance

$$\beta_{s,i}|\gamma_{s,i} \sim (1 - \gamma_{s,i}) \cdot \mathcal{N}(0, \tau^2) + \gamma_{s,i} \cdot \mathcal{N}(0, c^2\tau^2), \quad i = 1, \ldots, p.$$

The latent indicator variable $\gamma_{s,i}$ indicates the inclusion ($\gamma_{s,i} = 1$) or exclusion ($\gamma_{s,i} = 0$) of variable i in the model for subgroup s. We assume equal variances for all regression coefficients. We set the hyperparameters to the fixed values $\tau = 0.0375$ and $c = 20$ following Treppmann et al. [35]. This choice corresponds to a standard deviation of $c \cdot \tau = 0.75$ and a 95% probability interval of $[-1.47, 1.47]$ for $p(\beta_{s,i}|\gamma_{s,i} = 1)$.

Prior on γ linking variable and graph selection

The standard prior for the binary variable selection indicators $\gamma_{s,i}$ is a product of independent Bernoulli distributions as utilized by Treppmann et al. [35]. However, this does not consider information from other subgroups and relationships between covariates. For this situation, we propose a Markov random field (MRF) prior for the latent variable selection indicators that incorporates information on the relationships among the covariates as described by an undirected graph. This prior assumes that neighboring covariates in the graph are more likely to have a common effect and encourages their joint inclusion. The MRF prior for $\boldsymbol{\gamma}$ given $\boldsymbol{G}$ is defined as

$$p(\boldsymbol{\gamma}|\boldsymbol{G}) = \frac{\exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\{0,1\}^{pS}}\exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})} \propto \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma}),$$

where $\boldsymbol{\gamma} = (\gamma_{1,1}, \ldots, \gamma_{1,p}, \ldots, \gamma_{S,1}, \ldots, \gamma_{S,p})'$ is a pS-dimensional vector of variable inclusion indicators, $\boldsymbol{G}$ is a symmetric ($pS \times pS$) adjacency matrix representation of the graph, and a, b are scalar hyperparameters.

The hyperparameter a influences the overall variable inclusion probability and controls the sparsity of the model, with smaller values resulting in sparser models. Without loss of generality $a < 0$. The hyperparameter $b > 0$ determines the prior belief in the strength of relatedness between pairs of neighboring variables in the graph and controls

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 8 of 29

the probability of their joint inclusion. Higher values of *b* encourage the selection of variables with neighbors already selected into the model. The idea becomes more evident by looking at the conditional probability

$$p(\gamma_{s,i}|\boldsymbol{\gamma}_{-(s,i)}, \boldsymbol{G}) = \frac{\exp\left(a\gamma_{s,i} + 2b\gamma_{s,i} \cdot \left(\sum_{j\neq i} \gamma_{s,j} g_{ss,ij} + \sum_{r\neq s} \gamma_{r,i} g_{rs,ii}\right)\right)}{1 + \exp\left(a + 2b \cdot \left(\sum_{j\neq i} \gamma_{s,j} g_{ss,ij} + \sum_{r\neq s} \gamma_{r,i} g_{rs,ii}\right)\right)}.$$

An MRF prior for variable selection has also been used by other authors [21, 28, 33, 34]. However, unlike us, they do not address the problem of borrowing information across subgroups by linking covariates in a graph.

We propose a joint graph with possible edges between all pairs of covariates within each subgroup and edges between the same covariates in different subgroups. The elements $g_{rs,ij}$ in the adjacency matrix of the graph $\boldsymbol{G}$ represent the presence ($g_{rs,ij} = 1$) or absence ($g_{rs,ij} = 0$) of an edge between nodes (genes) i and j in subgroups r and s. They can be viewed as latent binary indicator variables for edge inclusion. The adjacency matrix in the present model is defined as

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} & \dots & \boldsymbol{G}_{1S} \\ \boldsymbol{G}_{12} & \boldsymbol{G}_{22} & \dots & \boldsymbol{G}_{2S} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{G}_{1S} & \boldsymbol{G}_{2S} & \dots & \boldsymbol{G}_{SS} \end{pmatrix}.$$

$\boldsymbol{G}_{ss} = (g_{ss,ij})_{i<j}$ is the matrix of latent edge inclusion indicators within subgroup s
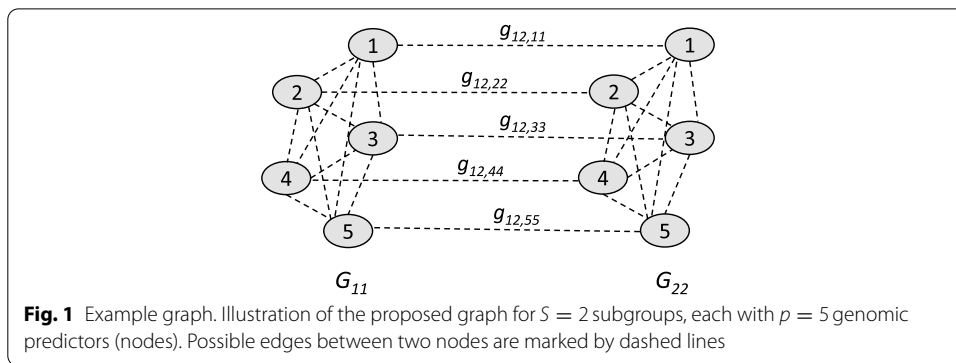
$$\boldsymbol{G}_{ss} = \begin{pmatrix} 0 & g_{ss,12} & \cdots & g_{ss,1(p-1)} & g_{ss,1p} \\ g_{ss,12} & 0 & \ddots & & g_{ss,2p} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ g_{ss,1(p-1)} & & \ddots & 0 & g_{ss,(p-1)p} \\ g_{ss,1p} & g_{ss,2p} & \cdots & g_{ss,(p-1)p} & 0 \end{pmatrix},$$

and $\boldsymbol{G}_{rs} = (g_{rs,ii})_{r<s}$ is the matrix of latent edge inclusion indicators between subgroups r and s

$$\boldsymbol{G}_{rs} = \mathrm{diag}(g_{rs,11}, \dots, g_{rs,pp}),$$

with $r, s = 1, \dots, S, r < s, i, j = 1, \dots, p, i < j$.

Thus, within each subgroup s we assume a standard undirected graph with possible edges between all pairs of genes representing conditional dependencies as in a functional or regulatory pathway. Between different subgroups we only allow for relations between the same gene in different subgroups (different genes in different subgroups are assumed to be unconnected). This allows sharing information between subgroups and prognostic genes shared by different subgroups have a higher inclusion probability. To visualize this idea, Fig. 1 shows an example network consisting of two subgroups, each with five predictors.

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 9 of 29

Fig. 1 Example graph. Illustration of the proposed graph for $S = 2$ subgroups, each with $p = 5$ genomic predictors (nodes). Possible edges between two nodes are marked by dashed lines

Graph selection prior on Ω and G

We infer the unknown graph and precision matrix using the structure learning approach for Gaussian graphical models by Wang [36]. The precision matrix of subgroup s corresponding to subgraph $\boldsymbol{G}_{ss}$ is denoted by $\boldsymbol{\Omega}_{ss} = (\omega_{ss,ij})_{i<j}$.

The corresponding prior is defined by

$$p(\boldsymbol{\Omega}_{ss}|\boldsymbol{G}_{ss}, \nu_0, \nu_1, \lambda) \propto \prod_{i<j} \mathcal{N}(\omega_{ss,ij}|0, \nu_{g_{ss,ij}}^2) \prod_i \text{Exp}(\omega_{ss,ii}|\frac{\lambda}{2}) \mathbb{1}_{\{\Omega_s \in \mathcal{M}^+\}},$$

with fixed hyperparameters $\nu_0 > 0$ small, $\nu_1 > 0$ large and $\lambda > 0$.

We assume the binary edge inclusion indicators within subgroup s ($g_{ss,ij}$) as well as between subgroups r and s ($g_{rs,ii}$) to be independent Bernoulli a priori

$$p(\boldsymbol{G}|\pi_G) \propto \prod_s \prod_{i<j} \left[\pi_G^{g_{ss,ij}} (1 - \pi_G)^{1-g_{ss,ij}}\right] \cdot \prod_{r<s} \prod_i \left[\pi_G^{g_{rs,ii}} (1 - \pi_G)^{1-g_{rs,ii}}\right],$$

with fixed prior probability of edge inclusion $\pi_G \in (0, 1)$.

Posterior inference

The joint posterior distribution for the set of all parameters $\theta = \{\boldsymbol{h}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{G}, \boldsymbol{\Omega}\}$ is proportional to the product of the joint likelihood and the prior distributions of the parameters in all subgroups

$$p(\boldsymbol{h}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{G}, \boldsymbol{\Omega}|\mathfrak{D}, \boldsymbol{X})$$

$$\propto \prod_{s=1}^{S} \left[L(\mathfrak{D}_s|\boldsymbol{\beta}_s, \boldsymbol{h}_s) \cdot p(\boldsymbol{X}_s|\boldsymbol{\Omega}_{ss})\right]$$

$$\cdot \prod_{s=1}^{S} \left[p(\boldsymbol{\Omega}_{ss}|\boldsymbol{G}_{ss}) \cdot p(\boldsymbol{G}) \cdot p(\boldsymbol{\gamma}|\boldsymbol{G}) \cdot \prod_{i=1}^{p} p(\beta_{s,i}|\gamma_{s,i}) \cdot \prod_{g=1}^{J_s} p(h_{s,g}|\boldsymbol{\beta}_s)\right].$$

Markov Chain Monte Carlo sampling

Markov Chain Monte Carlo (MCMC) simulations are required to obtain a posterior sample of the parameters. The different parameters are updated iteratively according

to their conditional posterior distributions using a Gibbs sampler. A brief outline of the MCMC sampling scheme is given in the following. More details are provided in the Appendix.

1 For subgroup $s = 1, \ldots, S$ update $\mathbf{\Omega}_{ss}$ with the block Gibbs sampler proposed by Wang [36].

2 Update all elements in $\boldsymbol{G}$ iteratively with Gibbs sampler from the conditional distributions $p(g_{ss,ij} = 1 | \boldsymbol{G}_{-ss,ij}, \omega_{ss,ij}, \boldsymbol{\gamma})$ as well as
$p(g_{rs,ii} = 1 | \boldsymbol{G}_{-rs,ii}, \boldsymbol{\gamma})$, where $\boldsymbol{G}_{-rs,ii}$ ($\boldsymbol{G}_{-ss,ij}$) denotes all elements in $\boldsymbol{G}$ except for $g_{rs,ii}$ ($g_{ss,ij}$).

3 Update all elements in $\boldsymbol{\gamma}$ iteratively with Gibbs sampler from the conditional distributions $p(\gamma_{s,i} = 1 | \boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}, \beta_{s,i})$, where $\boldsymbol{\gamma}_{-s,i}$ denotes all elements in $\boldsymbol{\gamma}$ except for $\gamma_{s,i}$.

4 Update $\beta_{s,i}$ from the conditional distribution $p(\beta_{s,i} | \boldsymbol{\beta}_{s,-i}, \boldsymbol{\gamma}_s, \boldsymbol{h}_s, \mathfrak{D}_s)$, $s = 1, \ldots, S$, $i = 1, \ldots, p$, using a random walk Metropolis-Hastings algorithm with adaptive jumping rule as proposed by Lee et al. [20]. $\boldsymbol{\beta}_{s,-i}$ includes all elements in $\boldsymbol{\beta}_s$ except for $\beta_{s,i}$.

5 The conditional distribution $p(h_{s,g} | \boldsymbol{h}_{s,-g}, \boldsymbol{\beta}_s, \boldsymbol{\gamma}_s, \mathfrak{D}_s)$ for the update of $h_{s,g}$ can be well approximated by the gamma distribution

$$h_{s,g} | \boldsymbol{h}_{s,-g}, \boldsymbol{\beta}_s, \boldsymbol{\gamma}_s, \mathfrak{D}_s \overset{\text{approx.}}{\sim} \mathcal{G}\left(a_0(H^*(c_{s,g}) - H^*(c_{s,g-1})) + d_{s,g}, a_0 + \sum_{k \in \mathcal{R}_{s,g} - \mathcal{D}_{s,g}} \exp(\boldsymbol{\beta}_{s'} \boldsymbol{x}_{s,k})\right),$$

where $d_{s,g}$ is the number of events in interval g for subgroup s and $\boldsymbol{h}_{s,-g}$ denotes the vector $\boldsymbol{h}_s$ without the g-th element, $g = 1, \ldots, J_s, s = 1, \ldots, S$ [17, chapter 3.2.2].

Starting with an arbitrary set of initial values for the parameters, the MCMC algorithm runs with a reasonably large number of iterations to obtain a representative sample from the posterior distribution. All subsequent results are based on single MCMC chains, each with 20 000 iterations in total and a burn-in period of 10 000 iterations. As starting values we choose an empty model with:

$$\boldsymbol{G}^{(0)} = \mathbf{0}_{pS \times pS}$$
$$\boldsymbol{\Sigma}_s^{(0)} = \boldsymbol{I}_{p \times p} \text{ and } \boldsymbol{\Omega}_{ss}^{(0)} = (\boldsymbol{\Sigma}_s^{(0)})^{-1} \text{for } s = 1, \ldots, S$$
$$\boldsymbol{\gamma}_s^{(0)} = (0, \ldots, 0)' \text{for } s = 1, \ldots, S$$
$$\beta_{s,i}^{(0)} \sim \mathcal{U}[-0.02, 0.02] \text{ for } i = 1, \ldots, p, s = 1, \ldots, S$$
$$h_{s,g}^{(0)} \sim \mathcal{G}(1, 1) \text{ for } s = 1, \ldots, S, g = 1, \ldots, J_s$$

We assessed convergence of each MCMC chain by looking at autocorrelations, trace plots and running mean plots of the regression coefficients. In addition, we ran several independent MCMC chains with different starting values to ensure that the chains and burn-in period were long enough to reach (approximate) convergence.

Posterior estimation and variable selection
We report the results of the Cox models in terms of marginal and conditional posterior means and standard deviations of the estimated regression coefficients, as well as posterior selection probabilities. After removal of the burn-in samples, the remaining MCMC samples serve as draws from the posterior distribution to calculate the empirical

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 11 of 29

estimates. These estimates are then averaged across all training sets for each variable separately.

The strategy for variable selection follows Treppmann et al. [35]. First, the mean model size m^* is computed as the average number of included variables across all MCMC iterations after the burn-in. Then the m^* variables with the highest posterior selection probability are considered as the most important variables and selected in the final model.

We visually assess the inferred graph by the marginal posterior probabilities of the pairwise edge inclusion indicators. High probabilities suggest that an edge exists between two covariates (nodes). We consider the presence of an edge as a continuous parameter rather than choosing a cutoff for binary decision.

Prediction

We use training data for model fitting and posterior estimation and test data to assess model performance. We evaluate the prediction performance of the Cox models by the integrated Brier score.

The expected Brier score can be interpreted as a mean square error of prediction. It measures the inaccuracy by comparing the estimated survival probability $\hat{S}(t|\boldsymbol{x}_m)$ of a patient $m, m = 1, .., n$, with the observed survival status $\mathbb{1}(\tilde{t}_m > t)$

$$\widehat{BS}(t) = \frac{1}{n} \sum_{m=1}^{n} \hat{w}_m(t) \cdot \left(\mathbb{1}(\tilde{t}_m > t) - \hat{S}(t|\boldsymbol{x}_m) \right)^2$$

and the squared residuals are weighted using inverse probability of censoring weights

$$\hat{w}_m(t) = \frac{\mathbb{1}(\tilde{t}_m \le t)\delta_m}{\hat{C}(\tilde{t}_m)} + \frac{\mathbb{1}(\tilde{t}_m > t)}{\hat{C}(t)}$$

to adjust for the bias caused by the presence of censoring in the data. $\hat{C}(t)$ is the Kaplan-Meier estimator of the censoring times [3, 31].

The predictive performance of competing survival models can be compared by plotting the Brier score over time (prediction error curves). Alternatively, prediction error curves can be summarized in one value with the integrated Brier score as a measure of inaccuracy over a time interval rather than at single time points [15]

$$IBS(t^*) = \frac{1}{t^*} \int_0^{t^*} BS(t)\mathrm{d}t, \quad t^* > 0.$$

Median probability model and Bayesian model averaging

For the calculation of the prediction error, we account for the uncertainty in model selection by two different approaches: the Median Probability Model (MPM) [1] and an approximation to Bayesian Model Averaging (BMA) [16]. After removal of the burn-in samples, we compute the Brier score over the "best" selected models. According to the BMA approach we choose the top 100 models with the largest log-likelihood values to obtain the marginal posterior means of the regression coefficients, which in turn are required for the risk score. Our choice of the number of top models for BMA

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 12 of 29

approximation is based on visual assessment of the MCMC frequencies of the different top-selected models. However, the number of models could be optimized. For the MPM approach we select all covariates with a mean posterior selection probability larger than 0.5. For these variables we calculate the marginal posterior means of the regression coefficients and the corresponding risk score.

Simulation study

In the following, we compare the performance of our proposed model, referred to as *CoxBVS-SL* (for Cox model with Bayesian Variable Selection and Structure Learning, as an extension of the model by Treppmann et al. [35]), to a standard subgroup model and a combined model. The combined model pools data from all subgroups and treats them as one homogeneous cohort, whereas the subgroup model only uses information in the subgroup of interest and ignores the other subgroups. Both standard approaches follow the Bayesian Cox model proposed by Treppmann et al. [35] with stochastic search variable selection and independent Bernoulli priors for the variable inclusion indicators γ.

The priors for variable selection and structure learning are specified as follows. We set the hyperparameter of the Bernoulli distribution to $\pi_\gamma = 0.02$, matching the prior probability of variable inclusion in the MRF prior of the CoxBVS-SL model. Based on a sensitivity analysis, we choose the hyperparameters of the MRF prior as $a = -4$ and $b = 1$. When the graph G contains no edges or $b = 0$ then the prior variable inclusion probability is $\frac{exp(a)}{(1+exp(a))} \approx 0.018$. This probability increases when $b > 0$ is combined with a non-empty graph.

For the sensitivity analysis of a and b we considered in total 36 combinations of the following hyperparameter values: $a \in \{-4, -3.75, -3.5, \ldots, -2.25, -2\}$ and $b \in \{0.25, 0.5, 0.75, 1\}$ and simulated the data according to scenario I with $n = p = 100$. Visual assessment of the results showed that they were relatively robust without major differences between the parameter combinations (Additional file 1: Fig. S1). Therefore, we selected the combination of values for a and b based on a compromise between variable selection accuracy (trade-off between large probability of true positive and small probability of false positive selections) and prediction performance.

The remaining hyperparameters for G and $\mathbf{\Omega}_{ss}$ are chosen as $\nu_0 = 0.1$, $\nu_1 = 10$, $\lambda = 1$ and $\pi_G = 2/(p-1)$, in accordance with Wang [36] and Peterson et al. [28]. Wang [36] extensively studied the impact of different parameter combinations on the structure learning results, reporting that the results were relatively insensitive to the choice of $\lambda = 1$. He recommended a range for the parameters ν_0 and ν_1 as providing good convergence and mixing. Based on his recommendation, we performed a sensitivity analysis in previous simulations to confirm that the parameter range is also appropriate for our experiments. All tested parameter combinations provided reasonable variable selection results with only small differences, which led us to choose one of the best performing combinations in terms of variable selection accuracy.

In the following simulations, we examine varying numbers of genomic covariates p and sample sizes n, with a focus on small sample sizes relative to the number of variables which is characteristic for gene expression data. We standardize the genomic covariates before model fitting and evaluation to have zero mean and unit variance. Parameters of the training data (mean and standard deviation of each variable) are used to scale the

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 13 of 29

training and test data. For the standard subgroup model and the proposed model we standardize each subgroup separately, whereas for the combined model we pool training data of all subgroups.

For Bayesian inference, typically one training data set is used for posterior estimation and an independent test data set for model evaluation. However, results have shown some variation due to the data draw. Therefore, in the following, simulation of training and test data is repeated ten times for each simulation scenario.

In a second simulation set up we use two different hyperparameters b for the subgraphs G_{ss}, $s = 1, 2$ and G_{12} in the MRF prior of the CoxBVS-SL model and compare the prediction performance with the *Sub-struct* model. In the latter G_{12} is an empty graph and only information of G_{ss} is included in the MRF prior. We use the same training and test data as before but only consider simulation scenarios with $p = 100$.

Data simulation

Training and test data, each consisting of n samples and p genomic covariates, are simulated from the same distribution as described in the following. We consider two subgroups that differ only in their relationship between genomic covariates and survival endpoint ($\boldsymbol{\beta}_s$, $s = 1, 2$), and in the parameters for the simulation of survival data. We generate gene expression data from the same multivariate normal distribution with mean vector $\mathbf{0}$ and precision matrix $\boldsymbol{\Omega}$. The precision matrix is defined such that the variance of each gene is 1 and partial correlations exist only between the first nine prognostic genes. Within the three blocks of prognostic genes determined by the same effect (gene 1 to 3, gene 4 to 6, and gene 7 to 9) we assume pairwise partial correlations of 0.5. All remaining genes are assumed to be uncorrelated.

We simulate survival data from a Weibull distribution according to Bender et al. [2], with scale η_s and shape κ_s parameters estimated from two gene expression cancer cohorts [4, 10] to obtain realistic survival times. Therefore, we compute survival probabilities at 3 and 5 years using the Kaplan-Meier estimator, separately for both cohorts. The corresponding probabilities are 57% and 75% for 3-years survival, and 42% and 62% for 5-years survival, respectively. Event times for subgroup s are simulated from

$$T_s \sim \left(-\frac{\log(U)}{\eta_s \exp(\boldsymbol{x}_s \boldsymbol{\beta}_s)} \right)^{1/\kappa_s}, \quad U \sim \mathcal{U}[0, 1],$$

with true effects $\boldsymbol{\beta}_s \in \mathbb{R}^p$, $s = 1, 2$. We randomly draw non-informative censoring times C_s from a Weibull distribution with the same Weibull parameters as for the event times, resulting in approximately 50% censoring rates in both subgroups. The individual observed event indicators and times until an event or censoring are defined as $\delta_s = \mathbb{1}(T_s \leq C_s)$ and $\widetilde{T}_s = \min(T_s, C_s)$, $s = 1, 2$.

We choose the true effects of the genomic covariates on survival outcome as stated in Table 1. Genes 1, 2, 3 and 7, 8, 9 are subgroup-specific, while genes 4, 5 and 6 have the same effect in both subgroups. All remaining genes represent noise and have no effect in both subgroups.

Simulation results I

We consider three low-dimensional settings with $p = 20$ genes and $n = 50, 75, 100$ samples in each subgroup, as well as five high-dimensional settings with $p = 100$ and sample sizes $n = 50, 75, 100, 150$. We also tested $p = 100$ and $n = 125$, but as expected, the results always lay between the results for $n = 100$ and $n = 150$. For this reason, they are not shown here. We compare our proposed model (*CoxBVS-SL*) to the standard subgroup model (*Subgroup*) and the standard combined or pooled model (*Pooled*) regarding variable selection accuracy and prediction performance.

Posterior selection probabilities for each gene are computed based on all iterations after the burn-in and averaged across all training data sets. The resulting mean posterior selection probabilities of the first nine genes in subgroup 1 are depicted in Fig. 2 (and in Additional file 1: Fig. S2, for subgroup 2). Across all simulation scenarios, the CoxBVS-SL model has more power for the selection of prognostic genes compared to the two standard approaches, and at the same time, does not erroneously select noise genes (false positives) as the Pooled model. As expected, with larger n, power and accuracy in variable selection increase for both, the CoxBVS-SL and the Subgroup model. The Pooled model only correctly identifies the joint effects of genes 4, 5 and 6 but fails to detect subgroup-specific effects.

Posterior estimates of the regression coefficients $\hat{\beta}_j$ of the first nine genes in subgroup 1 are shown in Fig. 3 for conditional posterior means (conditional on $\gamma = 1$) and in Additional file 1: Fig. S3 for marginal posterior means (independent of γ), both along with standard deviations. The corresponding results for subgroup 2 are depicted in Additional file 1: Figs. S4 and S5. For $n < 100$ the conditional posterior means of the prognostic genes are less shrunk than the marginal posterior means. Results of the CoxBVS-SL model and the Subgroup model are very similar, whereas the Pooled model averages
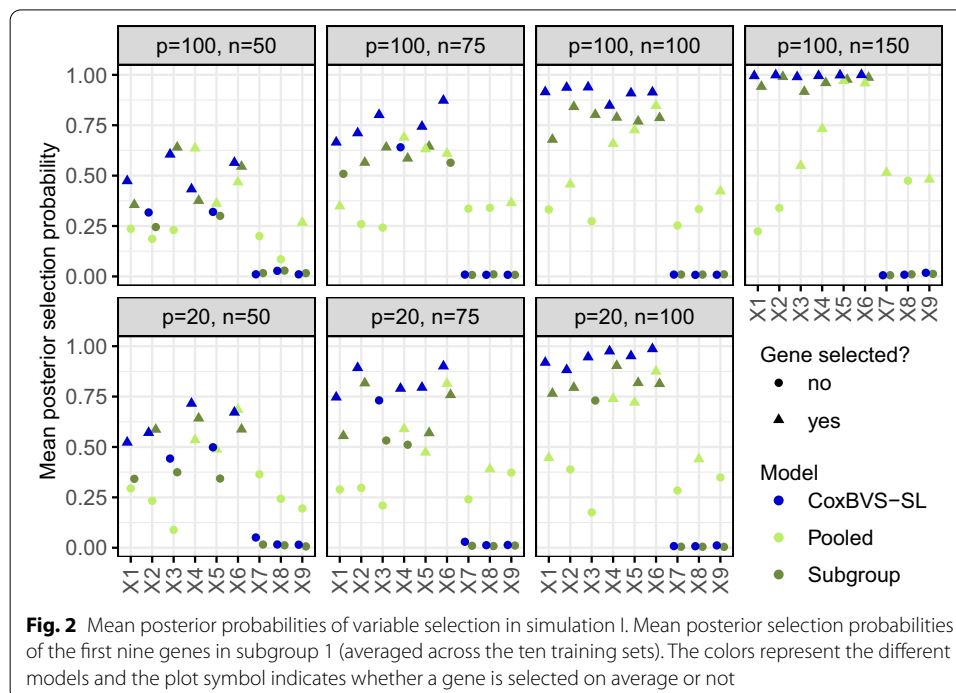

Fig. 2 Mean posterior probabilities of variable selection in simulation I. Mean posterior selection probabilities of the first nine genes in subgroup 1 (averaged across the ten training sets). The colors represent the different models and the plot symbol indicates whether a gene is selected on average or not
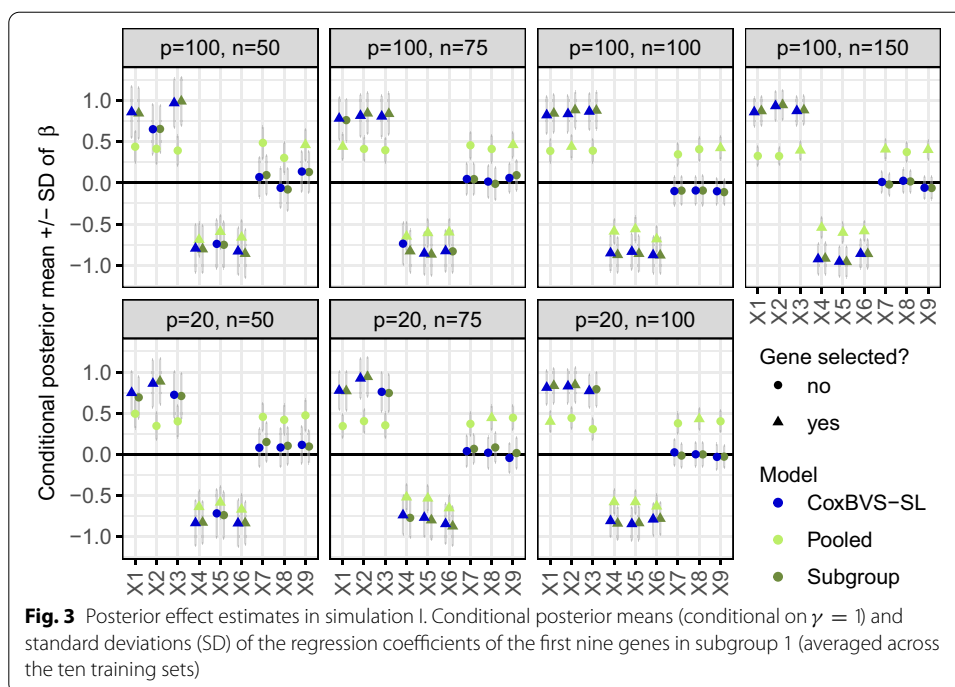
Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 15 of 29

Fig. 3 Posterior effect estimates in simulation I. Conditional posterior means (conditional on $\gamma = 1$) and standard deviations (SD) of the regression coefficients of the first nine genes in subgroup 1 (averaged across the ten training sets)

Table 1 Effects in simulation I

	Gene											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	$\cdots$	**p**
$\boldsymbol{\beta}_1$	1	1	1	−1	−1	−1	0	0	0	0	$\cdots$	0
$\boldsymbol{\beta}_2$	0	0	0	−1	−1	−1	1	1	1	0	$\cdots$	0

True effects in both subgroups for the simulation of survival outcome

effects across subgroups leading to biased subgroup-specific effects and more false positives. Surprisingly, the joint effects of genes 4, 5 and 6 are also more precisely estimated (less shrunk) by CoxBVS-SL and Subgroup compared to Pooled.

We assess prediction performance by the integrated Brier score (IBS), computed based on the Median Probability Model (MPM, Fig. 4 for subgroup 1 and Additional file 1: Fig. S7, for subgroup 2) and the Bayesian Model Averaging (BMA, Additional file 1: Fig. S6, for subgroup 1 and Additional file 1: Fig. S8, for subgroup 2). The Pooled model has the worst prediction accuracy. In the case of MPM, CoxBVS-SL performs clearly better than Subgroup, for BMA both models are competitive.

Inference of the graph showed relatively high accuracy for learning the conditional dependence structure among genes within subgroups and for detecting joint effects across different subgroups. The block correlation structure between the prognostic genes within each subgroup is correctly estimated by the precision matrix and the subgraph $\boldsymbol{G}_{ss}$, $s = 1, 2$ in the CoxBVS-SL model (see Additional file 1: Fig. S9). Inference of the subgraph $\boldsymbol{G}_{12}$ linking both subgroups improves with increasing sample size. The corresponding marginal posterior edge inclusion probabilities of the prognostic genes with joint effects (genes 4, 5 and 6) are larger than for the remaining genes, which

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 16 of 29

becomes more evident for increasing n (see Additional file 1: Fig. S10). Findings support the assumption that incorporating network information into variable selection may increase power to detect associations with the survival outcome and improve prediction accuracy.

Simulation results II

Next, we study the effect of two different hyperparameters b in the MRF prior of the CoxBVS-SL model with respect to variable selection and prediction performance. The new hyperparameter $b_1 = 1$ corresponds to the subgraphs $G_{ss}, s = 1, 2$ within each subgroup and $b_2 = 1, 1.5, 2, 2.5, 3$ to the subgraph G_{12} linking both subgroups. By choosing a larger value for b_2, we give G_{12} more weight in the MRF prior and thus, increase the prior variable inclusion probability for genes being simultaneously selected in both subgroups and having a link in G_{12}.

We compare the results of CoxBVS-SL with varying b_2 to the results of the *Sub-struct* model where $b_2 = 0$ and only information of $G_{ss}, s = 1, 2$ is included in the MRF prior. In this comparison we investigate how much information is added by G_{12} over G_{ss}. For the other hyperparameters we use the same values as in the previous simulations. We apply all models to the same training and test data sets as before but only consider simulation scenarios with $p = 100$ and $n = 50, 75, 100, 125, 150$.

Figure 5 shows the mean posterior selection probabilities of the first nine genes in subgroup 1 (subgroup 2 is presented in Additional file 1: Fig. S11). The results of Sub-struct are similar to CoxBVS-SL with $b_2 = 1$. Increasing values of b_2 lead to larger posterior variable inclusion probabilities, however, not only for the prognostic genes (see genes 7, 8 and 9 in subgroup 1). This means more power for the correct identification of prognostic genes when $n \leq p$, but on the other hand, a tendency towards more false positives.

Posterior estimates of the regression coefficients $\hat{\beta}_j$ are very similar for all models. Figure 6 shows the conditional posterior means (conditional on $\gamma = 1$) and Additional file 1: Fig. S12 the marginal posterior means (independent of γ) along with standard deviations of the first nine genes in subgroup 1. The corresponding results of subgroup 2 are depicted in Additional file 1: Figs. S13 and S14.
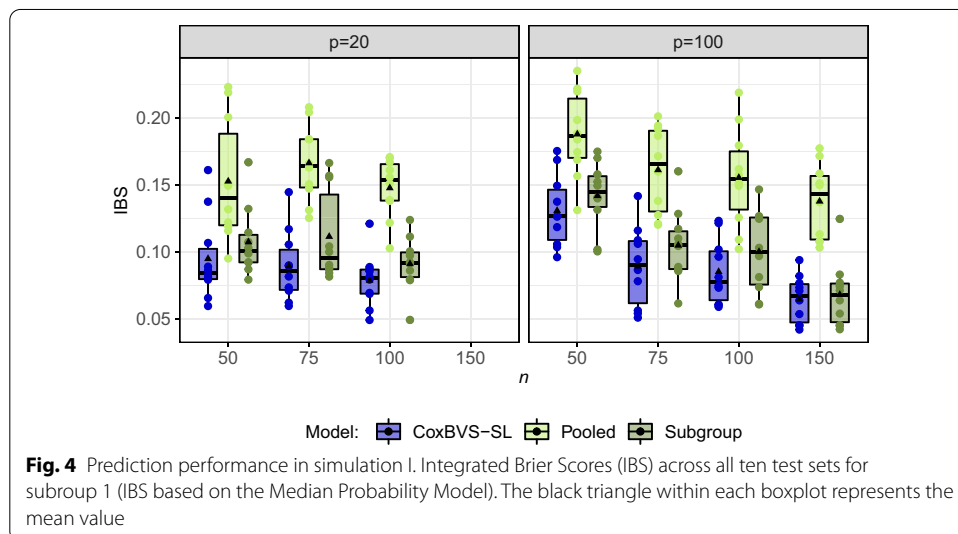
We assess prediction performance in terms of the integrated Brier score (IBS), computed based on the Median Probability Model (Fig. 7) and the Bayesian Model Averaging (Additional file 1: Fig. S15). Larger values of b_2 tend to lead to a slightly better prediction performance of CoxBVS-SL compared to Sub-struct when $n < p$. When the sample size is large, the prediction accuracy of all models is similarly good.

Additional file 1: Fig. S16 compares the results of the subgraph G_{12} for varying b_2 in CoxBVS-SL. For larger values of b_2 the marginal posterior edge inclusion probabilities of the prognostic genes with joint effects (genes 4, 5 and 6) increase, as expected, since they are given a higher weight in the prior. However, when $b_2 = 3$ we also notice a minor increase of the marginal posterior edge inclusion probabilities of the other six prognostic genes with subgroup-specific effects.
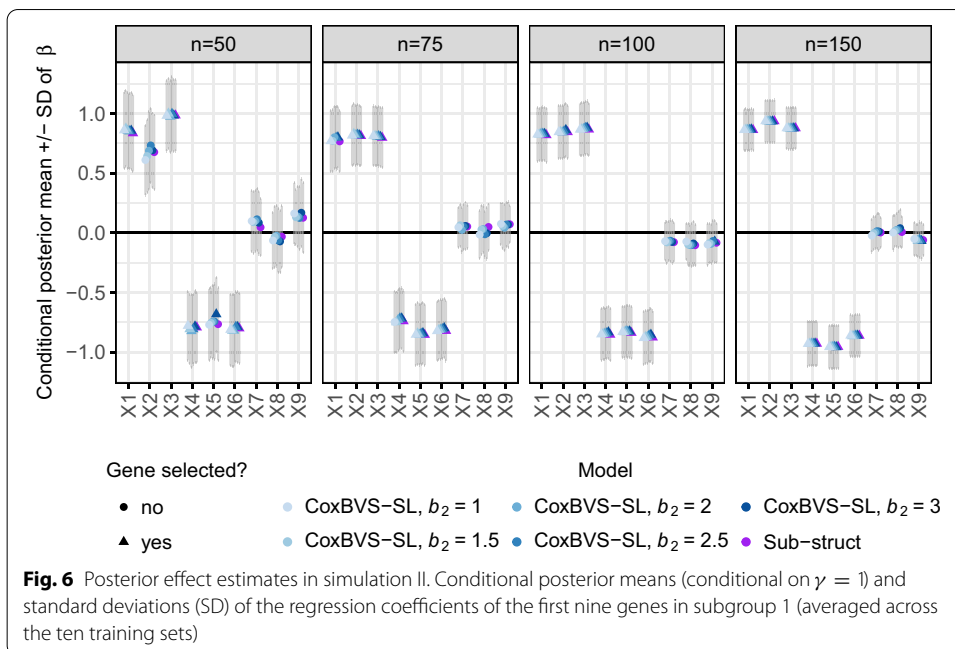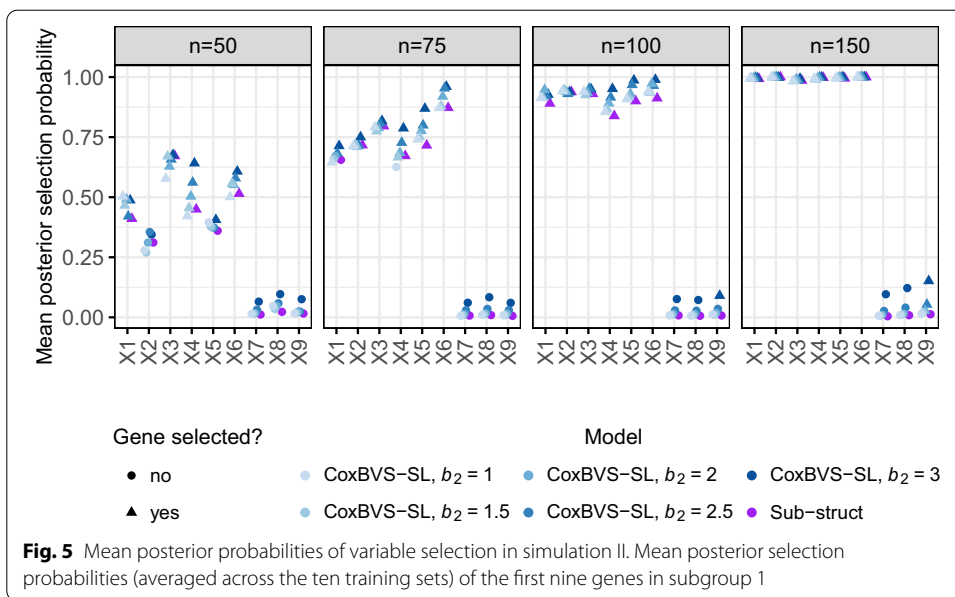
Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 17 of 29

Table 2 Effects in simulation II

Protein	β_1	β_2
Akt	2	0
Akt_pS473	2	0
Akt_pT308	2	0
EGFR	0	2
EGFR_pY1068	0	2
EGFR_pY1173	0	2
AMPK_alpha	−1.5	1.5
Annexin.1	1.5	−1.5
GSK3.alpha.beta	−2	−2
GSK3.alpha.beta_pS21_S9	−2	−2
GSK3_pS9	−2	−2
X14.3.3_beta	0	0
X14.3.3_epsilon	0	0
X14.3.3_zeta	0	0
X4E.BP1	0	0
X4E.BP1_pS65	0	0
X4E.BP1_pT37T46	0	0
X4E.BP1_pT70	0	0
X53BP1	0	0
A.Raf_pS299	0	0

Simulated effects in both subgroups. Groups of proteins with the same effect are defined by different phosphorylation sites (or isoforms) of the same protein so that they can learn from each other
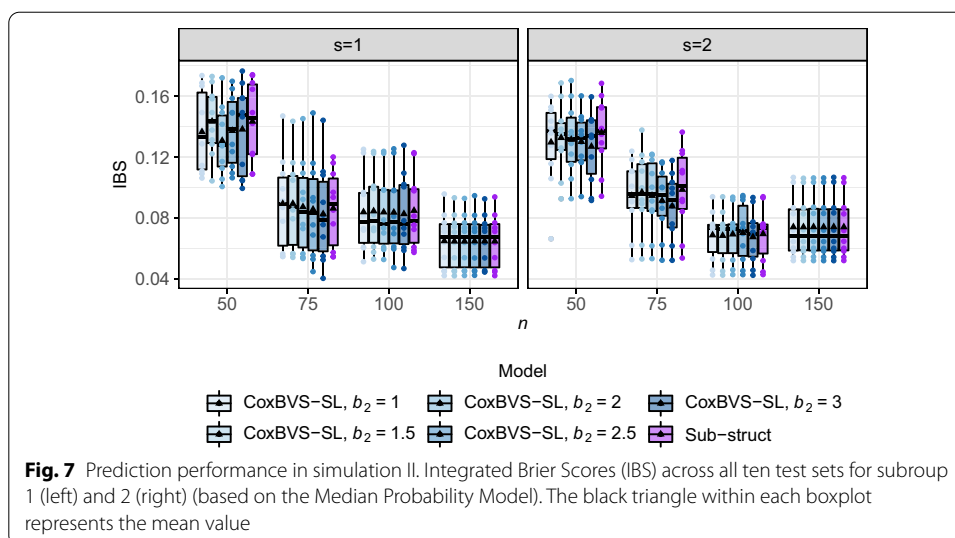

Fig. 4 Prediction performance in simulation I. Integrated Brier Scores (IBS) across all ten test sets for subroup 1 (IBS based on the Median Probability Model). The black triangle within each boxplot represents the mean value

Case study based on Glioblastoma protein expression data

In this section we compare CoxBVS-SL with varying b_2 to both standard models, Pooled and Subgroup. We use the Glioblastoma protein expression data from Peterson et al. [28], comprising 212 samples with survival data (159 events) and $p = 187$ proteins. For reasons of computation time, we use only $p = 20$ proteins and standardize the protein expression data as described in the previous section. In contrast to the

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 18 of 29

Fig. 5 Mean posterior probabilities of variable selection in simulation ll. Mean posterior selection probabilities (averaged across the ten training sets) of the first nine genes in subgroup 1

Fig. 6 Posterior effect estimates in simulation ll. Conditional posterior means (conditional on $\gamma = 1$) and standard deviations (SD) of the regression coefficients of the first nine genes in subgroup 1 (averaged across the ten training sets)

previous simulations, we do not draw the expression data from a multivariate normal distribution, but instead use real protein expression data with realistic correlation structure between all covariates, following the concept of plasmode simulations as described by Franklin et al. [12]. We still simulate the relationship between proteins and survival outcome by choosing artificial effects and simulating the survival data from a Weibull distribution. We randomly divide the complete data set into two equally large subsets to obtain two subgroups.
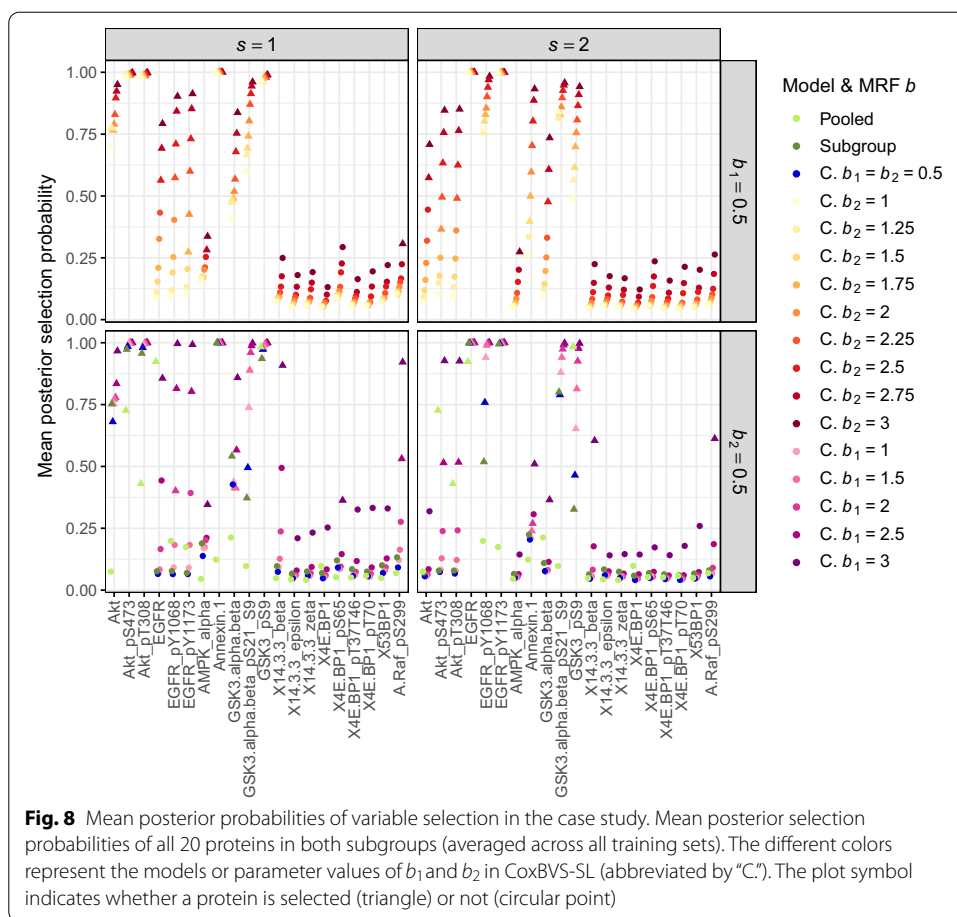
Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 19 of 29

Fig. 7 Prediction performance in simulation II. Integrated Brier Scores (IBS) across all ten test sets for subroup 1 (left) and 2 (right) (based on the Median Probability Model). The black triangle within each boxplot represents the mean value

For the survival endpoint we simulate the event times T_s and censoring times C_s, respectively, in subgroup s from a Weibull distribution with scale and shape parameters estimated by the Kaplan-Meier estimator of the true event and censoring times, respectively, in the specific subgroup. The individual observed event indicators and survival times until an event or censoring are defined as $\delta_s = \mathbb{1}(T_s \leq C_s)$ and $t_s = \min(T_s, C_s)$, resulting in approximately 42% censoring rates in both subgroups. The effects in subgroup $s = 1$ and $s = 2$ that we assume for the simulation of survival data are depicted in Table 2.

We repeatedly randomly split the complete data into training (with proportion 0.8) and test sets, stratified by subgroup and event indicator. In total, we generate ten training data sets for model fitting and ten test data sets for evaluation of the prediction performance.

We choose the hyperparameters in accordance with the case study in Peterson et al. [28] as follows. For the two standard models a prior probability of variable inclusion of 0.2 is assumed. In the CoxBVS-SL model we set the hyperparameters of the precision matrix and graph to $\nu_0 = 0.6, \nu_1 = 360, \lambda = 1$ and $\pi_G = 2/(p - 1)$. The hyperparameters of the MRF prior are $a = -1.75, b = 0.5$ and as in the previous section, we tried out two different values for b: $b_1 = 0.5$ and $b_2 = 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3$, or $b_1 = 1, 1.5, 2, 2.5, 3$ and $b_2 = 0.5$.
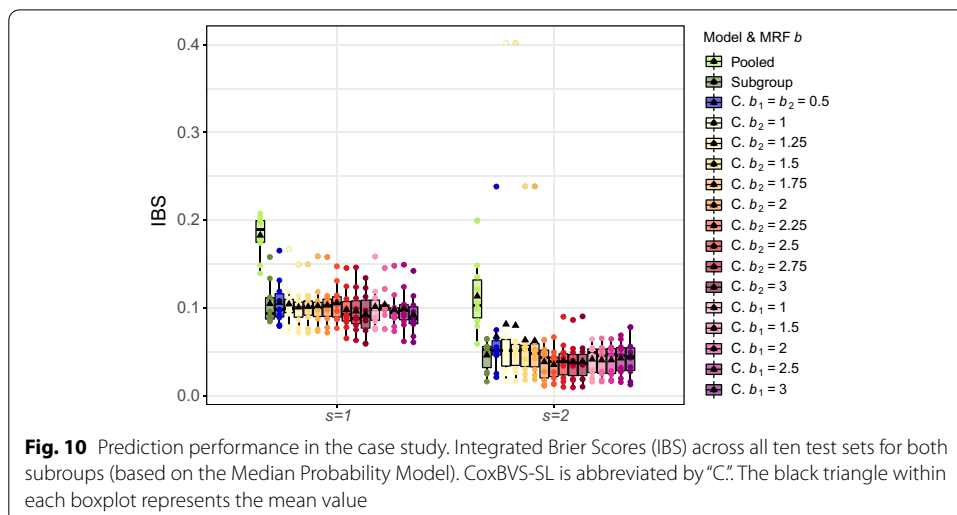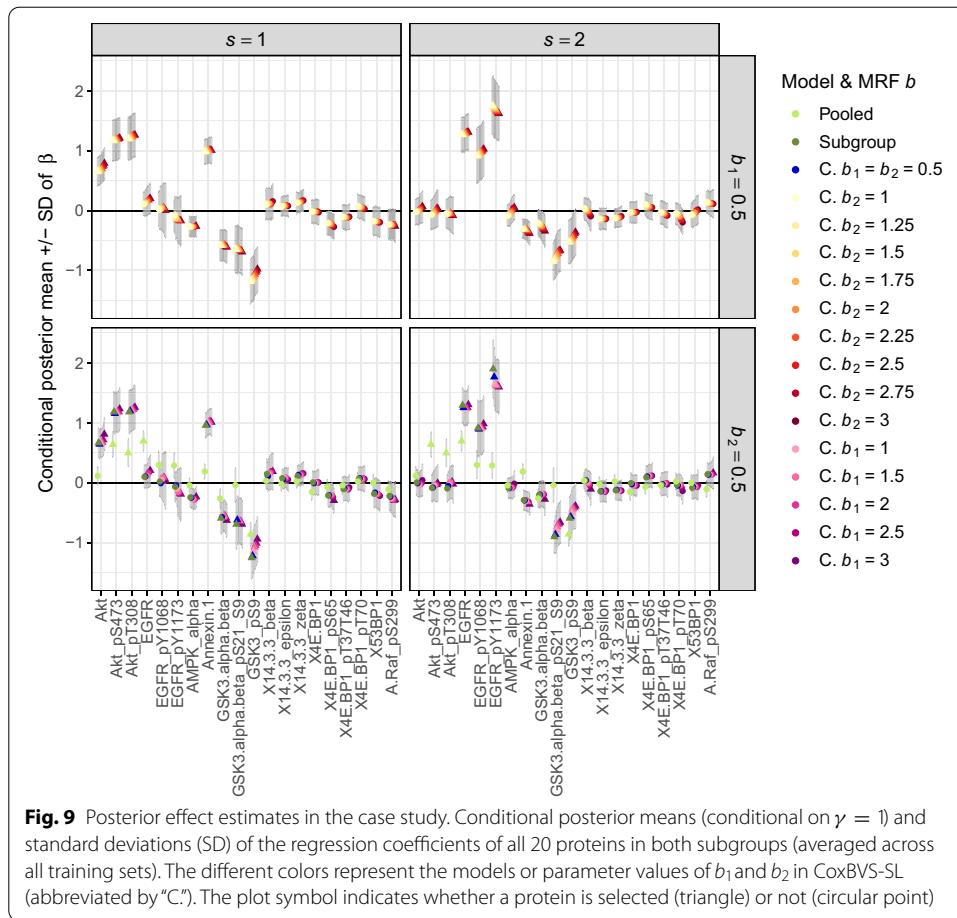
Results of the case study

When either b_1 or b_2 increases the mean posterior selection probabilities of all proteins increase too (Fig. 8). The Subgroup and CoxBVS-SL model with $b_1 = b_2 = 0.5$ perform similarly. They correctly identify the subgroup-specific effects of the first six proteins and do not falsely select any noise proteins. Interestingly, the effects of proteins AMPK and Annexin (ID 7 and 8), going in opposite directions for both subgroups, as well as the joint effects of proteins GSK3 are not all identified. There are a few false negatives. The Pooled model, in contrast, shows a clear bias for the subgroup-specific and opposite effects. The effects are averaged across both subgroups,

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 20 of 29

Fig. 8 Mean posterior probabilities of variable selection in the case study. Mean posterior selection probabilities of all 20 proteins in both subgroups (averaged across all training sets). The different colors represent the models or parameter values of b_1 and b_2 in CoxBVS-SL (abbreviated by "C."). The plot symbol indicates whether a protein is selected (triangle) or not (circular point)

which also becomes evident when looking at the posterior estimates of the regression coefficients, for the conditional posterior means in Fig. 9 and for the marginal posterior means in Additional file 1: Fig. S17. The results of the Subgroup and CoxBVS-SL model are similar. In particular, the posterior means of the noise proteins with null effect are close to 0, also for large values of b_1 or b_2.

When we compare all models with regard to prediction accuracy in Fig. 10 and Additional file 1: Fig. S18, we again see competitive performance for the Subgroup and CoxBVS-SL model whereas Pooled is clearly worse. We can observe a tendency towards slightly improved prediction accuracy for increasing values of b_2.

Finally, we assess the impact of increasing values of b_2 on the subgraph G_{12} linking both subgroups. The corresponding marginal posterior edge selection probabilities are depicted in Additional file 1: Fig. S19. When b_2 becomes larger first, the posterior edge selection probabilities of proteins 8, 10 and 11 with opposite or joint effects in both subgroups increase, followed by the first six proteins with subgroup-specific effects and protein 9 with joint effect. The posterior edge selection probabilities of the noise proteins in both subgroups remain at the prior mean and only start to increase slightly when $b_2 \geq 2.5$. Proteins 7 and 9 have much smaller posterior edge selection probabilities than the other proteins with opposite or joint effects, which fits to previous findings.

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 21 of 29

Fig. 9 Posterior effect estimates in the case study. Conditional posterior means (conditional on $\gamma = 1$) and standard deviations (SD) of the regression coefficients of all 20 proteins in both subgroups (averaged across all training sets). The different colors represent the models or parameter values of b_1 and b_2 in CoxBVS-SL (abbreviated by "C."). The plot symbol indicates whether a protein is selected (triangle) or not (circular point)

Fig. 10 Prediction performance in the case study. Integrated Brier Scores (IBS) across all ten test sets for both subroups (based on the Median Probability Model). CoxBVS-SL is abbreviated by "C.". The black triangle within each boxplot represents the mean value

When b_1 becomes larger, the marginal posterior edge selection probabilities in the subgraphs G_{11} and G_{22} show no visible changes. In G_{12} they increase for some proteins however, to a much lesser extent than for larger b_2.

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 22 of 29

Discussion

We consider the situation of different, possibly heterogeneous patients subgroups (preknown cohorts or data sets) with survival endpoint and continuous molecular measurements such as gene expression data. When building a separate risk prediction model for each subgroup, it is important to consider heterogeneity but at the same time it can be reasonable to allow sharing information across subgroups to increase power, in particular when the sample sizes are small. For this situation we propose a hierarchical Cox model with stochastic search variable selection prior. To achieve higher power in variable selection and better prediction performance, we use an MRF prior instead of the standard Bernoulli prior for the latent variable selection indicators γ. The MRF prior leads to higher selection probabilities for genes that are related in an undirected graph. We use this graph to link genes across different subgroups and thereby borrow information between subgroups. Genes that are simultaneously prognostic in different subgroups have a higher probability of being selected into the respective subgroup Cox models. As a side aspect, the graph in the MRF prior also allows us to estimate a network between genes within each subgroup providing indications of functionally related genes and pathways. Here, genes that are conditionally dependent have a higher selection probability.

In the simulations and the case study we compared our proposed CoxBVS-SL model to the standard approach with independent Bernoulli prior for γ represented by the Subgroup and Pooled model. Simulations showed that the Pooled model performed worst in terms of variable selection and prediction accuracy. It averaged the effects across both subgroups and thus, led to biased estimates. CoxBVS-SL had more power in variable selection and slightly better prediction performance than Subgroup when the sample size was small. For $n > p$ both models were competitive. However, in the case study the CoxBVS-SL and Subgroup model performed similarly well (Pooled was again clearly worse). The reason for this may be that the sample sizes in both subgroups were relatively large, in particular $n > p$.

In the MRF prior of our proposed CoxBVS-SL model we specify one unique hyperparameter b for both, the connection levels of covariates within and between subgroups. Since this assumption may be inadequate, we considered further simulations where we studied the effect of increasing values of b_2, representing the weight that is given to the subgraph $\boldsymbol{G}_{12}$ in the MRF prior of CoxBVS-SL, and compared the results to the Substruct model where $b_2 = 0$. When b_2 was small, CoxBVS-SL and Sub-struct performed very similarly. Thus, the subgraph linking both subgroups had only a small influence on the results compared to the conditional dependencies among covariates within each subgroup (subgraphs $\boldsymbol{G}_{11}$ and $\boldsymbol{G}_{22}$). For larger values of b_2 prediction performance slightly improved and power in variable selection increased but on the other hand, there was a tendency towards false positive variables. By using different hyperparameters b_1 and b_2, we can vary the strength of connection between pairs of covariates within and between subgroups. However, we still assume that all pairs of subgroups are equally-likely linked a priori. In our situation this assumption is justified since we have no prior knowledge of the amount of shared, similar effects between subgroups. If prior information on the heterogeneity structure between subgroups (similar effects) is available, it can be incorporated into the MRF prior or the graph prior.

The problem of different connection levels of covariates within and between subgroups can in a similar way be approached by the hyperparameter π_G in the graph prior, instead of the hyperparameter b in the MRF prior. In previous simulations (data not shown) we increased the weight for $\boldsymbol{G}_{12}$ by choosing a larger value for the prior probability of edge inclusion π_G for the corresponding edge inclusion indicators $g_{12,ii}$, $i = 1, \ldots, p$. This led to larger posterior edge selection probabilities, however, for all genes and not only the ones with joint effects. The variable selection results did not change remarkably. We could observe a small increase in power for all genes which again implied a tendency towards false positives.

Due to computation time, we have included only up to 200 variables so far and the analysis of many thousands of genes is not (yet) feasible. An advantage of the CoxBVS-SL model is that it does not require prior knowledge of the graph among the covariates within and between subgroups. It accounts for uncertainty over both variable and graph selection. In situations where pathway information is available and the graph structure is known, it is possible to incorporate this structural information in the MRF prior via a fixed graph.

We assume that subgroups are pre-specified with the subgroup affiliation of each patient being unique and fixed. However, in situations with unknown subgroups the latent subgroup structure would first need to be determined using methods such as clustering. A wide variety of approaches have been proposed for the clustering of molecular data such as gene expression profiles [9, 13, 18, 25, 39] with extensions to sparse clustering [32, 38] and integrative clustering of multiple omics data types [6, 19].

Conclusions

To our knowledge, we propose the first completely Bayesian approach to combine different, possibly heterogeneous subgroups/cohorts in Cox regression with variable selection. We offer a solution for sharing information across the subgroups to increase power in variable selection and improve prediction performance.

We were able to demonstrate the superiority of our proposed CoxBVS-SL model over the two standard approaches. The standard Pooled model always performed worst, whereas the CoxBVS-SL model outperformed the standard Subgroup model when $n \leq p$ and otherwise was competitive. This suggests that incorporating network information into variable selection can increase power to identify the prognostic covariates and improve prediction performance. We showed that a proper choice of the hyperparameter b (and a) in the MRF prior is crucial for the results of the graph and the Cox model.

Our proposed model does not require prior knowledge of the dependency structure between covariates within subgroups and the heterogeneity structure between subgroups (i.e., of the amount of shared, similar effects). In the absence of any prior structural information, we assume that all pairs of covariates within and between subgroups are equally-likely linked a priori, and we allow inference of the corresponding unknown graphical structures. In situations where prior structural information is available, for example pathway information or degree of heterogeneity between subgroups, this information can be incorporated into our model. We presented a way to assign different connection levels to covariates within and between subgroups by using different hyperparameters b in the MRF

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 24 of 29

prior. Alternatively, one could use fixed edges in the graph or varying prior edge selection probabilities.

The discovery of graphical structure is an additional benefit of our proposed model. However, our focus is on prediction performance, unbiased effect estimation and variable selection in the Cox model. Our proposed CoxBVS-SL model showed improved results in the situation of small sample sizes which is an important problem, not only in clinical applications.

Appendix

Details of the MCMC algorithm

In the following, steps 1 to 4 of the MCMC sampling scheme in the Methods section are explained in more detail.

Step 1: Update of Ω_{ss}

The block Gibbs sampler proposed by Wang [36] is used to update Ω_{ss} for subgroups $s = 1, \ldots, S$. The conditional distribution of Ω_{ss} is

$$
\begin{aligned}
p(\Omega_{ss}|G_{ss}, X_s) & \\
\propto\ & p(X_s|\Omega_{ss}) \cdot p(\Omega_{ss}|G_{ss}) \\
\propto\ & |\Omega_{ss}|^{n_s/2} \exp\{-\frac{1}{2}\mathrm{tr}(S_s\Omega_{ss})\} \cdot \prod_{i<j} \exp\{-\frac{1}{2}\frac{\omega_{ss,ij}^2}{v_{g_{ss,ij}}^2}\} \cdot \prod_i \exp\{-\frac{\lambda}{2}\omega_{ss,ii}\}.
\end{aligned}
$$

Consider the following partitions

$$
\Omega_{ss} = \left(\begin{array}{c|c} \widetilde{\Omega}_{11} & \widetilde{\omega}_{12} \\ \hline \widetilde{\omega}_{12}' & \widetilde{\omega}_{22} \end{array} \right) = \left(\begin{array}{cccc|c} \omega_{ss,11} & \omega_{ss,12} & \cdots & \omega_{ss,1(p-1)} & \omega_{ss,1p} \\ \omega_{ss,12} & \omega_{ss,22} & \cdots & \omega_{ss,2(p-1)} & \omega_{ss,2p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \omega_{ss,1(p-1)} & \omega_{ss,2(p-1)} & \cdots & \omega_{ss,(p-1)(p-1)} & \omega_{ss,(p-1)p} \\ \hline \omega_{ss,1p} & \omega_{ss,2p} & \cdots & \omega_{ss,(p-1)p} & \omega_{ss,pp} \end{array} \right)
$$

and analogously

$$
S_s = X_s'X_s = \left(\begin{array}{c|c} \widetilde{S}_{11} & \widetilde{s}_{12} \\ \hline \widetilde{s}_{12}' & \widetilde{s}_{22} \end{array} \right), \quad V_s = (v_{g_{ss,ij}}^2) = \left(\begin{array}{c|c} \widetilde{V}_{11} & \widetilde{v}_{12} \\ \hline \widetilde{v}_{12}' & 0 \end{array} \right),
$$

where V_s is a $(p \times p)$ symmetric matrix with zeros on the diagonal. For the block update of Ω_{ss} focus on the last column (and row) of Ω_{ss}: $(\widetilde{\omega}_{12}, \widetilde{\omega}_{22})$ with $\widetilde{\omega}_{12} = (\omega_{ss,1p}, \omega_{ss,2p}, \ldots, \omega_{ss,(p-1)p})'$, $\widetilde{\omega}_{22} = \omega_{ss,pp}$.

The conditional distribution of the last column of Ω_{ss} is

$$
\begin{aligned}
p(\widetilde{\omega}_{12}, \widetilde{\omega}_{22}|X_s, G_{ss}, \widetilde{\Omega}_{11}) \propto\ & \left(\widetilde{\omega}_{22} - \widetilde{\omega}_{12}'\widetilde{\Omega}_{11}^{-1}\widetilde{\omega}_{12}\right)^{n_s/2} \\
& \cdot \exp\left\{ -\frac{1}{2}\left[\widetilde{\omega}_{12}'\mathrm{diag}(\widetilde{v}_{12}^{-1})\widetilde{\omega}_{12} + 2\widetilde{s}_{12}'\widetilde{\omega}_{12} + (\widetilde{s}_{22} + \lambda)\widetilde{\omega}_{22}\right] \right\}.
\end{aligned}
$$

Consider the following transformations

$$
u = \widetilde{\omega}_{12}, \quad v = \widetilde{\omega}_{22} - \widetilde{\omega}_{12}'\widetilde{\Omega}_{11}^{-1}\widetilde{\omega}_{12}.
$$

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 25 of 29

Then the conditional distribution is

$$
p(\boldsymbol{u}, v | \boldsymbol{X}_s, \boldsymbol{G}_{ss}, \widetilde{\boldsymbol{\Omega}}_{11}) \propto \underbrace{v^{n_s/2} \exp\left\{ -\frac{\widetilde{s}_{22} + \lambda}{2} v \right\}}_{(*_1)}
$$

$$
\cdot \underbrace{\exp\left\{ -\frac{1}{2} \left[\boldsymbol{u}' \underbrace{\left(\mathrm{diag}(\widetilde{\boldsymbol{v}}_{12}^{-1}) + (\widetilde{s}_{22} + \lambda) \widetilde{\boldsymbol{\Omega}}_{11}^{-1} \right)}_{= \boldsymbol{C}^{-1}} \boldsymbol{u} + 2\widetilde{\boldsymbol{s}}_{12}' \boldsymbol{u} \right] \right\}}_{(*_2)}
$$

$(*_1) \propto \mathcal{G}(v | \frac{n_s}{2} + 1, \frac{\widetilde{s}_{22}+\lambda}{2})$,

$(*_2) \propto \mathcal{N}(\boldsymbol{u} | - \boldsymbol{C}\widetilde{\boldsymbol{s}}_{12}, \boldsymbol{C})$.

Permuting any column in $\boldsymbol{\Omega}_{ss}$ to be updated to the last one leads to a block Gibbs sampler for the update of $\boldsymbol{\Omega}_{ss}$.

Step 2: Update of *G*

Update all elements in $\boldsymbol{G}$ iteratively with Gibbs sampler from their conditional distributions. All elements $g_{rs,ij}$ are assumed independent Bernoulli a priori with $p(g_{rs,ij} = 1) = \pi_G$ and $p(g_{rs,ij} = 0) = 1 - \pi_G$.

Update $g_{rs,ii}$, $r, s = 1, \ldots, S$, $r < s$, $i = 1, \ldots, p$ (edges between the same gene in different subgroups) from the conditional distribution

$$
p(g_{rs,ii} | \boldsymbol{G}_{-rs,ii}, \boldsymbol{\gamma}) = \frac{p(g_{rs,ii}) \cdot p(\boldsymbol{\gamma} | \boldsymbol{G}_{-rs,ii}, g_{rs,ii})}{\sum_{g_{rs,ii} \in \{0,1\}} p(g_{rs,ii}) \cdot p(\boldsymbol{\gamma} | \boldsymbol{G}_{-rs,ii}, g_{rs,ii})},
$$

where $\boldsymbol{G}_{-rs,ii}$ denotes all elements in $\boldsymbol{G}$ except for $g_{rs,ii}$. Accept $g_{rs,ii} = 1$ with probability

$$
p(g_{rs,ii} = 1 | \boldsymbol{G}_{-rs,ii}, \boldsymbol{\gamma}) = \frac{w_a}{w_a + w_b},
$$

where

$$
w_a = \pi_G \cdot \exp(a \mathbf{1}_{pS}' \boldsymbol{\gamma} + b \boldsymbol{\gamma}' \boldsymbol{G} \boldsymbol{\gamma})|_{g_{rs,ii}=1}
$$
$$
w_b = (1 - \pi_G) \cdot \exp(a \mathbf{1}_{pS}' \boldsymbol{\gamma} + b \boldsymbol{\gamma}' \boldsymbol{G} \boldsymbol{\gamma})|_{g_{rs,ii}=0} .
$$

This means, update $g_{rs,ii}$ as follows: $g_{rs,ii} = \begin{cases} 1, & \text{if } u < \frac{w_a}{w_a+w_b}, \ u \sim \mathcal{U}[0,1] \\ 0, & \text{else}. \end{cases}$

Update $g_{ss,ij}$, $s = 1, \ldots, S$, $i, j = 1, \ldots, p$, $i < j$ (edges between different genes in the same subgroup) from the conditional distribution

$$
p(g_{ss,ij} | \boldsymbol{G}_{-ss,ij}, \omega_{ss,ij}, \boldsymbol{\gamma}) = \frac{p(g_{ss,ij}) \cdot p(\omega_{ss,ij}, \boldsymbol{\gamma} | \boldsymbol{G}_{-ss,ij}, g_{ss,ij})}{\sum_{g_{ss,ij} \in \{0,1\}} p(g_{ss,ij}) \cdot p(\omega_{ss,ij}, \boldsymbol{\gamma} | \boldsymbol{G}_{-ss,ij}, g_{ss,ij})}
$$
$$
\propto p(g_{ss,ij}) \cdot p(\omega_{ss,ij} | g_{ss,ij}) \cdot p(\boldsymbol{\gamma} | \boldsymbol{G}_{-ss,ij}, g_{ss,ij}),
$$

where $\boldsymbol{G}_{-ss,ij}$ denotes all elements in $\boldsymbol{G}$ except for $g_{-ss,ij}$. Accept $g_{ss,ij} = 1$ with probability

$$
p(g_{ss,ij} = 1 | \boldsymbol{G}_{-ss,ij}, \omega_{ss,ij}, \boldsymbol{\gamma}) = \frac{w_a}{w_a + w_b},
$$

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 26 of 29

where

$$w_a = \pi_G \cdot \mathcal{N}(\omega_{ss,ij}|0, v_1^2) \cdot \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{g_{ss,ij}=1}$$
$$w_b = (1 - \pi_G) \cdot \mathcal{N}(\omega_{ss,ij}|0, v_0^2) \cdot \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{g_{ss,ij}=0}.$$

Step 3: Update of γ

Update $\gamma_{s,i}$, $s = 1, \ldots, S$, $i = 1, \ldots, p$, with Gibbs sampler from the conditional distribution

$$
\begin{aligned}
p(\gamma_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}, \beta_{s,i}) &= \frac{p(\gamma_{s,i}, \beta_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})}{\sum_{\gamma_{s,i}\in\{0,1\}} p(\gamma_{s,i}, \beta_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})} \\
&= \frac{p(\gamma_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})}{\sum_{\gamma_{s,i}\in\{0,1\}} p(\gamma_{s,i}|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}, \boldsymbol{G})} \\
&= \frac{p(\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}|\boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i})}{\sum_{\gamma_{s,i}\in\{0,1\}} p(\gamma_{s,i}, \boldsymbol{\gamma}_{-s,i}|\boldsymbol{G}) \cdot p(\beta_{s,i}|\gamma_{s,i})},
\end{aligned}
$$

where $\boldsymbol{\gamma}_{-s,i}$ denotes all elements in $\boldsymbol{\gamma}$ except for $\gamma_{s,i}$. Accept $\gamma_{s,i} = 1$ with probability

$$p(\gamma_{s,i} = 1|\boldsymbol{\gamma}_{-s,i}, \boldsymbol{G}, \beta_{s,i}) = \frac{w_a}{w_a + w_b},$$

where

$$w_a = \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{\gamma_{s,i}=1} \cdot \mathcal{N}(\beta_{s,i}|0, c^2\tau^2)$$
$$w_b = \exp(a\mathbf{1}'_{pS}\boldsymbol{\gamma} + b\boldsymbol{\gamma}'\boldsymbol{G}\boldsymbol{\gamma})|_{\gamma_{s,i}=0} \cdot \mathcal{N}(\beta_{s,i}|0, \tau^2).$$

Step 4: Update of β

A random walk Metropolis-Hastings algorithm with adaptive jumping rule as proposed by Lee et al. [20] is used to update $\beta_{s,i}$ for $s = 1, \ldots, S$ and $i = 1, \ldots, p$. The full conditional posterior distribution of $\beta_{s,i}$ is

$$
\begin{aligned}
&p(\beta_{s,i}|\boldsymbol{\beta}_{s,-i}, \boldsymbol{\gamma}_s, \boldsymbol{h}_s, \mathfrak{D}_s) \\
&\propto L(\mathfrak{D}_s|\boldsymbol{\beta}_s, \boldsymbol{h}_s) \cdot p(\boldsymbol{\beta}_s|\boldsymbol{\gamma}_s) \\
&\propto \prod_{g=1}^{J_s} \left[\exp(-h_{s,g} \sum_{k\in\mathcal{R}_{s,g}-\mathcal{D}_{s,g}} \exp(\boldsymbol{\beta}'_s\boldsymbol{x}_{s,k})) \prod_{l\in\mathcal{D}_{s,g}} [1 - \exp(-h_{s,g}\exp(\boldsymbol{\beta}'_s\boldsymbol{x}_{s,l}))] \right] \\
&\cdot \exp(-\frac{1}{2}\boldsymbol{\beta}'_s\Sigma^{-1}_{\beta_s}\boldsymbol{\beta}_s),
\end{aligned}
$$

where $\boldsymbol{\beta}_{s,-i}$ denotes the vector $\boldsymbol{\beta}_s$ without the i-th element.

$\Sigma_{\beta_s} = \text{diag}(\sigma^2_{\beta_{s,1}}, \ldots, \sigma^2_{\beta_{s,p}})$ with $\sigma^2_{\beta_{s,i}} = (1 - \gamma_{s,i}) \cdot \tau^2 + \gamma_{s,i} \cdot c^2\tau^2$.

In MCMC iteration t update $\beta_{s,i}$ as follows:

(i) Sample a proposal $\beta_{s,i}^{(prop)}$ from a proposal distribution
$q(\beta_{s,i}^{(prop)}|\beta_{s,i}^{(t-1)}) = \mathcal{N}(\beta_{s,i}^{(prop)}|\mu_{\beta_{s,i}}^{(t-1)}, v_{\beta_{s,i}}^{(t-1)})$

(ii) Calculate the ratio of ratios

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 27 of 29

$$r_{s,i} = \frac{p(\beta_{s,i}^{(prop)}|\boldsymbol{\beta}_{s,-i}^{(t-1)}, \boldsymbol{\gamma}_s^{(t-1)}, \boldsymbol{h}_s^{(t-1)}, \mathfrak{D}_s)/q(\beta_{s,i}^{(prop)}|\beta_{s,i}^{(t-1)})}{p(\beta_{s,i}^{(t-1)}|\boldsymbol{\beta}_{s,-i}^{(t-1)}, \boldsymbol{\gamma}_s^{(t-1)}, \boldsymbol{h}_s^{(t-1)}, \mathfrak{D}_s)/q(\beta_{s,i}^{(t-1)}|\beta_{s,i}^{(prop)})}$$

(iii) Accept the proposal $\beta_{s,i}^{(prop)}$ if $\min\{r_{s,i}, 1\} > u$ with $u \sim \mathcal{U}[0,1]$.

The mean and variance of the proposal distribution can be approximated based on the first and second derivative of the log conditional posterior distribution with respect to $\beta_{s,i}^{(t-1)}$.

Abbreviations
BMA: Bayesian model averaging; CoxBVS-SL: Cox model with Bayesian variable selection and structure learning; IBS: Integrated brier score; MCMC: Markov Chain Monte Carlo; MPM: Median probability model; MRF: Markov random field; SSVS: Stochastic search variable selection.

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04483-z.

> **Additional file 1**: Additional supporting information for the results of the simulation studies and the case study.

Declarations

Ethics approval and consent to participate
Only published data that are publicly available online were used in this paper.

Consent for publication
Not applicable

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany. [2] Department of Biostatistics, Oslo Centre for Biostatistics and Epidemiology, University of Oslo, 0317 Oslo, Norway.

References
1. Barbieri MM, Berger JO. Optimal predictive model selection. Ann Stat. 2004;32(3):870–97. https://doi.org/10.1214/009053604000000238.

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 28 of 29

2. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Stat Med. 2005;24(11):1713–23. https://doi.org/10.1002/sim.2059.

3. Binder H, Porzelius C, Schumacher M. An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. Biom J. 2011;53(2):170–89. https://doi.org/10.1002/bimj.201000152.

4. Botling J, Edlund K, Lohr M, Hellwig B, Holmberg L, Lambe M, Berglund A, Ekman S, Bergqvist M, Pontén F, König A, Fernandes O, Karlsson M, Helenius G, Karlsson C, Rahnenführer J, Hengstler JG, Micke P. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. Clin Cancer Res. 2013;19(1):194–204. https://doi.org/10.1158/1078-0432.CCR-12-1139.

5. Chakraborty S, Lozano AC. A graph Laplacian prior for Bayesian variable selection and grouping. Comput Stat Data Anal. 2019;136(C):72–91. https://doi.org/10.1016/j.csda.2019.01.00.

6. Chalise P, Koestler DC, Bimali M, Yu Q, Fridley BL. Integrative clustering methods for high-dimensional molecular data. Transl Cancer Res. 2014;3(3):202–16. https://doi.org/10.3978/j.issn.2218-676X.2014.06.03.

7. Cox DR. Regression models and life-tables. J Roy Stat Soc Ser B (Methodol). 1972;34(2):187–220.

8. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Ser B Stat Methodol. 2014;76(2):373–97. https://doi.org/10.1111/rssb.12033.

9. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinform. 2008;9:497. https://doi.org/10.1186/1471-2105-9-497.

10. Der SD, Sykes J, Pintilie M, Zhu C-Q, Strumpf D, Liu N, Jurisica I, Shepherd FA, Tsao M-S. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J Thorac Oncol. 2014;9(1):59–64. https://doi.org/10.1097/JTO.0000000000000042.

11. Drton M, Maathuis MH. Structure learning in graphical modeling. Annu Rev Stat Appl. 2017;4(1):365–93. https://doi.org/10.1146/annurev-statistics-060116-053803.

12. Franklin JM, Schneeweiss S, Polinski JM, Rassen JA. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. Comput Stat Data Anal. 2014;72:219–26. https://doi.org/10.1016/j.csda.2013.10.018.

13. Gao C, Zhu Y, Shen X, Pan W. Estimation of multiple networks in Gaussian mixture models. Electron J Stat. 2016;10(1):1133–54. https://doi.org/10.1214/16-EJS1135.

14. George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Stat Assoc. 1993;88(423):881–9. https://doi.org/10.1080/01621459.1993.10476353.

15. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. Stat Med. 1999;18(17–18):2529–45.

16. Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Stat Sci. 1999;14(4):382–401.

17. Ibrahim JG, Chen M-H, Sinha D (2005) Bayesian Survival Analysis, Corr. 2nd print. New York [u.a.]: Springer Series in Statistics. Springer.

18. Jaskowiak PA, Campello RJ, Costa IG. On the selection of appropriate distances for gene expression data clustering. BMC Bioinform. 2014;15(2):2. https://doi.org/10.1186/1471-2105-15-S2-S2.

19. Kirk P, Griffin JE, Savage RS, Ghahramani Z, Wild DL. Bayesian correlated clustering to integrate multiple datasets. Bioinformatics. 2012;28(24):3290–7. https://doi.org/10.1093/bioinformatics/bts595.

20. Lee KH, Chakraborty S, Sun J. Bayesian variable selection in semiparametric proportional hazards model for high dimensional survival data. Int J Biostat. 2011;7(1):1–32. https://doi.org/10.2202/1557-4679.1301.

21. Li F, Zhang NR. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. J Am Stat Assoc. 2010;105(491):1202–14. https://doi.org/10.1198/jasa.2010.tm08177.

22. Madjar K. Survival models with selection of genomic covariates in heterogeneous cancer studies. Dissertation, Faculty of Statistics, TU Dortmund University (2018). https://doi.org/10.17877/DE290R-19140

23. Madjar K, Rahnenführer J. Weighted cox regression for the prediction of heterogeneous patient subgroups. arXiv: 2003.08965 (2020)

24. Mitra R, Müller P, Ji Y. Bayesian graphical models for differential pathways. Bayesian Anal. 2016;11(1):99–124. https://doi.org/10.1214/14-BA931.

25. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, Achas M, Adebiyi E. Clustering algorithms: their application to gene expression data. Bioinform Biol Insights. 2016;10:38316. https://doi.org/10.4137/BBI.S38316.

26. Park T, Casella G. The Bayesian Lasso. J Am Stat Assoc. 2008;103(482):681–6. https://doi.org/10.1198/016214508000000337.

27. Peterson C, Stingo FC, Vannucci M. Bayesian inference of multiple Gaussian graphical models. J Am Stat Assoc. 2015;110(509):159–74. https://doi.org/10.1080/01621459.2014.896806.

28. Peterson CB, Stingo FC, Vannucci M. Joint Bayesian variable and graph selection for regression models with network-structured predictors. Stat Med. 2016;35(7):1017–31. https://doi.org/10.1002/sim.6792.

29. Richter J, Madjar K, Rahnenführer J. Model-based optimization of subgroup weights for survival analysis. Bioinformatics. 2019;35(14):484–91. https://doi.org/10.1093/bioinformatics/btz361.

30. Saegusa T, Shojaie A. Joint estimation of precision matrices in heterogeneous populations. Electron J Stat. 2016;10(1):1341–92. https://doi.org/10.1214/16-EJS1137.

31. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. Bioinformatics. 2007;23(14):1768–74. https://doi.org/10.1093/bioinformatics/btm232.

32. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009;25(22):2906–12. https://doi.org/10.1093/bioinformatics/btp543.

33. Stingo FC, Vannucci M. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. Bioinformatics. 2011;27(4):495–501. https://doi.org/10.1093/bioinformatics/btq690.

Madjar *et al. BMC Bioinformatics* 2021, **22**(1):586

Page 29 of 29

34. Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. Ann Appl Stat. 2011;5(3):1978–2002. https://doi.org/10.1214/11-AOAS463.

35. Treppmann T, Ickstadt K, Zucknick M. Integration of multiple genomic data sources in a Bayesian cox model for variable selection and prediction. Computational and Mathematical Methods in Medicine **vol. 2017 Article ID 7340565**, 2017;19. https://doi.org/10.1155/2017/7340565

36. Wang H. Scaling It Up: Stochastic Search Structure Learning in Graphical Models. Bayesian Anal. 2015;10(2):351–77. https://doi.org/10.1214/14-BA916.

37. Weyer V, Binder H. A weighting approach for judging the effect of patient strata on high-dimensional risk prediction signatures. BMC Bioinform. 2015;16:294. https://doi.org/10.1186/s12859-015-0716-8.

38. Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc. 2010;105(490):713–26. https://doi.org/10.1198/jasa.2010.tm09415.

39. Wiwie C, Baumbach J, Röttger R. Comparing the performance of biomedical clustering methods. Nat Methods. 2015;12(11):1033–8. https://doi.org/10.1038/nmeth.3583.

40. Yajima M, Telesca D, Ji Y, Muller P. Differential patterns of interaction and Gaussian graphical models. Collection of Biostatistics Research Archive, COBRA Preprint Series. 2012;91.

41. Zucknick M, Saadati M, Benner A. Nonidentical twins: comparison of frequentist and Bayesian lasso for Cox models. Biom J. 2015;57(6):959–81. https://doi.org/10.1002/bimj.201400160.

Publisher's Note