Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival

Journal:	*Statistical Methods in Medical Research*
Manuscript ID:	SMM-14-0020.R1
Manuscript Type:	Original Article
Keywords:	Survival Analysis, Bayesian Variable Selection, Reversible Jump, MCMC, Penalised Regression, Stability Selection, Breast Cancer, Gene Expression

Abstract:

As data-rich medical datasets are becoming routinely collected, there is a grow- ing demand for regression methodology that facilitates variable selection over a large number of predictors. Bayesian variable selection algorithms offer an attractive solution, whereby a sparsity inducing prior allows inclusion of sets of predictors simultaneously, leading to adjusted effect estimates and inference of which covariates are most important.

We present a new implementation of Bayesian variable selection, based on a Reversible Jump MCMC algorithm, for survival analysis under the Weibull regression model. A realistic simulation study is presented comparing against an alternative LASSO based variable selection strategy in datasets of up to 20,000 covariates. Across half the scenarios our new method achieved identical sensitivity and specificity to the LASSO strategy, and a marginal improvement otherwise. Runtimes were comparable for both approaches, taking approximately a day for 20,000 covariates. Subsequently, we present a real data application in which 119 protein-based markers are explored for association with breast cancer survival in a case cohort of 2,287 patients with ER-positive disease. Evidence was found for three independent prognostic tumour markers of survival, one of which is novel. Our new approach demonstrated the best specificity.

Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival

P. J . Newcombe[1], H. Raza Ali[2,3,4], F. M. Blows[5], E. Provenzano[6], P. D. Pharoah[4,5,7], C. Caldas[2,4,5], and S. Richardson[1]

[1]MRC Biostatistics Unit, Cambridge, UK

[2]Cancer Research UK Cambridge Institute

[3]Department of Pathology, University of Cambridge

[4]Cambridge Experimental Cancer Medicine Centre and NIHR Cambridge Biomedical Research Centre

[5]Department of Oncology, University of Cambridge

[6]NIH Cambridge Biomedical Research Centre

[7]Strangeways Research Laboratory

Abstract

As data-rich medical datasets are becoming routinely collected, there is a growing demand for regression methodology that facilitates variable selection over a large number of predictors. Bayesian variable selection algorithms offer an attractive solution, whereby a sparsity inducing prior allows inclusion of sets of predictors simultaneously, leading to adjusted effect estimates and inference of which covariates are most important.

We present a new implementation of Bayesian variable selection, based on a Reversible Jump MCMC algorithm, for survival analysis under the Weibull regression model. A realistic simulation study is presented comparing against an alternative LASSO based variable selection strategy in datasets of up to 20,000 covariates. Across half the scenarios our new method achieved identical sensitivity and specificity to the LASSO strategy, and a marginal improvement otherwise. Runtimes were comparable for both approaches, taking approximately a day for 20,000 covariates. Subsequently, we present a real data application in which 119 protein-based markers are explored for association with breast cancer survival in a case cohort of 2,287 patients with ER-positive disease. Evidence was found for three independent prognostic tumour markers of survival, one of which is novel. Our new approach demonstrated the best specificity.

Keywords: Survival analysis; Bayesian variable selection; Reversible Jump; Stability Selection; Breast Cancer; Gene Expression; Penalised regression; MCMC.

1 Introduction

As large data-rich studies are becoming routinely collected in medical research, there is a growing need for regression techniques designed to cope with many predictors. While the simplest approach is to analyse each variable one at a time, the results are difficult to interpret since confounding from between-predictor associations can cloud the location of true signals leading to elevated false positive rates. Ideally, when predictors are correlated, multivariate regression should be performed to account for the association structure and enable accurate inference on the subset of variables most likely to represent true effects. However, when the number of covariates is high, traditional Ordinary Least Squares methods suffer from over-fitting — the limited information available is spread too thinly among the covariates leading to unstable parameter estimates with high standard errors.

This inspired the development of LASSO penalised regression by Tibshirani in 1996[1] whereby a penalty term is included in the likelihood to encourage sparsity. The penalty term modifies the likelihood of the regression coefficients, with a large penalty leading to the exclusion of many variables. Typically the penalty is tuned through cross-validation such that covariates with negligible predictive effects are removed. The over-fitting problem is thus avoided and prediction improved. Over the years there have been a number of extensions to the original method, including: SCAD[2], Elastic Net[3], Adaptive LASSO[4] and Fused LASSO[5] each generating a class of penalties to address specific predictive aims. Some of these methods have been applied in the genomic context to explore multi-SNP models of disease[6, 7] or to search for master predictors[8]. Extensions to model structured sparsity via the group LASSO[9] or to impose additional hierarchical constraints, e.g. when searching for interactions[10] have been proposed. Techniques have been developed to obtain significance measures for the covariates, including resampling procedures[11, 12] and, recently, a formal significance test[13, 14], as well as a modified bootstrap procedure that provides a valid approximation to the LASSO distribution thereby enabling construction of uncertainty intervals[15].

An alternative to penalised regression is Bayesian sparse regression, in which posterior inference is made on the predictors, and subsets of predictors, most likely associated with outcome. Attractive features of Bayesian sparse regression include inference of posterior probabilities for each predictor, posterior inference on the model space and, perhaps most importantly, the possibility of natural incorporation of prior information into the analysis. A variety of formulations and methods for implementing Bayesian variable selection have been developed. George and McCulloch first proposed inducing sparsity via two-component 'spike and slab' mixture priors on the effect of each covariate, consisting of a 'spike' either exactly at or around zero, corresponding to exclusion from the model, and a flat 'slab' elsewhere[16]. Binary indicator variables are used to denote which component each covariate belongs to; the posterior expectation of which provides marginal posterior probabilities of effect. Sparsity is encouraged by placing priors on these indicators which favour the 'spike'. Such models are typically fitted using MCMC and a number of algorithms have been developed, varying in how the spike and slab components are formulated[16, 17, 18]. Notably, an adaptive shrinkage approach proposed by Hoti and Sillanpaa eases the computational challenge through use of single component normal priors, with a hyperprior on the precision that leads to an approximate spike and slab shape, thereby avoiding the use of indicator variables and mixture component switching. A cut-off on the magnitude of effect is used to define whether or not a covariate is included in the model[19].

Whereas these models implement variable selection through priors on each covariate, an alternative approach is to consider the model space as a whole and place priors on the number of included covariates. In 1995, Green demonstrated how classical MCMC methodology can be extended to explore models of differing dimensions using a 'Reversible Jump' algorithm in which the Metropolis-Hastings acceptance ratio is modified to account for addition and deletion of covariates during model updates[20]. The level of sparsity is controlled through a prior on the number of included covariates. Reversible Jump has been applied to model selection problems in many areas, including genomics and in particular genetic association analysis[21], meta-analysis[22, 23] and predictive model building[24, 25] in which the ability to incorporate prior information has been exploited in various ways.

A drawback of Reversible Jump, however, is that the dimension switching leads to a substantial increase in algorithmic complexity. In the case of linear regression, conjugate closed form expressions under the normal likelihood can be exploited to avoid MCMC sampling of covariate effects, allowing the stochastic search algorithm to focus exclusively on the model space, dramatically simplifying the mixing of the algorithm[26]. The 'Stochastic Shotgun Search' (SSS) algorithm utilises this principle, and in addition proposes a modified search algorithm which parallelises the exploration of potentially vast model spaces while focusing on areas of high posterior mass[27]. This allows rapid identification of models with high posterior mass, at the cost of 'formal' posterior inference since the model search space is deliberately restricted. Alternatively, the 'Evolutionary Stochastic Search' (ESS) algorithm, developed by Bottolo and Richardson, similarly utilises conjugate normality to integrate over covariate effects but allows exploration of the entire model space resulting in formal posterior inference on covariate and model probabilities[28]. Sophisticated and efficient implementations of ESS now exist for the analysis of continuous univariate and multivariate outcomes[29, 30]. These procedures are very fast and are capable of analysing thousands of predictors simultaneously. Superior power and specificity in comparison to penalised regression style approaches has been shown, which has facilitated the identification of novel genomic associations[30, 31]. For a more detailed overview of approaches to Bayesian Variable Selection we refer readers to the excellent review by O'Hara and Sillanpaa[32].

The Cox semi-parametric proportional hazards model is the most widely used approach for the analysis of right censored survival data. Cox regression is semi-parametric in that the baseline hazard is ascribed no particular form and is estimated non-parametrically. Working in the Bayesian framework, however, it was natural to choose a fully parametric survival model for the analysis we present in this paper. Whereas a proportional hazards model assumes that covariates multiply the hazard by some constant, so-called 'accelerated failure time' models are a class of (typically fully parametric) survival models in which the covariates are assumed to multiply the expected survival time. Consequently, regression parameter estimates from accelerated failure time models are more robust to omitted covariates[33]. The Weibull distribution is an appealing choice for fully parametric survival modelling since, uniquely, it has both the accelerated failure time and the proportional hazards property; there is a direct correspondence between the parameters under the two models[34]. Therefore hazard ratios can be inferred as in Cox regression, but while benefiting from the accelerated failure time property. In comparison to Cox regression, when the baseline hazard function describes the data well the Weibull model offers greater precision in the estimation of hazard ratios. Conversely, however, the non-parametric nature of the baseline hazard under a Cox model affords robustness over a wider range of survival trajectories.

Unfortunately, in the context of Weibull regression for survival analysis, there are no conjugate results to exploit and so we resort to Reversible Jump MCMC, sampling both regression parameters and models. This is, to our knowledge, the first application of a Reversible Jump algorithm to the Weibull model for survival analysis. After exploring performance in comparison to an alternative frequentist variable selection strategy (penalised Cox regression with stability selection), we present a real data application to explore tumour markers of breast cancer survival in a prospective case cohort. Further details of this study and dataset are given below.

2 Data

Breast cancer remains a significant public health problem with more than 45,000 cases diagnosed in the UK in 2012 and, despite significant improvements over the last thirty years[35, 36], continues to be a major cause of mortality amongst women in the western world. Treatment currently consists of surgical excision of the tumour and adjuvant therapies which may include radiotherapy, endocrine therapy, cytotoxic chemotherapy and targeted biological therapies depending on tumour characteristics and patient preference. However, there is substantial heterogeneity in patient response to these therapies, all of which are associated with significant toxicity. There is now a well established set of pathological prognostic factors for Breast Cancer including tumour size and grade, lymph node status, oestrogen receptor (ER) status and Human Epithermal growth factor Receptor 2 (HER2) status[37, 38] which are widely used in clinical practice to guide treatment decisions. For example, a patient with excellent prognosis may want to avoid exposing themselves to highly toxic therapies. However, our ability to reliably identify patients who can safely forgo adjuvant chemotherapy is limited impairing optimal clinical decision making. Breast cancer is now known to consist of a variety of molecular subtypes[39] and while these tools are of profound clinical utility, there is much scope to expand on this set of prognostic risk factors which do not currently reflect the whole variety of breast cancer leading to suboptimal clinical decisions, particularly the over-prescription of adjuvant chemotherapy[40].

We explore a large collection of predominantly protein-based markers related to cancer biology including markers of cancer stem cells and the tumour microenvironment, which may underpin the molecular diversity of tumours[41, 39, 42]. Our analysis is performed using cases from the ongoing population-based breast cancer cohort of the SEARCH (studies of epidemiology and risk factors in cancer heredity) study; a genetic epidemiology study with a molecular pathology component recruiting individuals resident in the east of England. Ascertainment of breast cancer cases was conducted by the East Anglia Cancer Registry. The study includes both prevalent and incident cases. Prevalent cases are those who were already diagnosed with breast cancer at the time of study commencement. Specifically, these included women diagnosed with invasive breast cancer under the age of 55 between 1991 and mid-1996 and still alive in 1996. Incident cases are those individuals diagnosed after study commencement. These were women under the age of 70 at the time of breast cancer diagnosis after mid-1996. The two different ER subtypes of breast cancer (positive and negative) are recognised as markedly different diseases biologically and pathologically with demonstrated differences in baseline hazard over time[43]. Therefore it is sensible to stratify on this characteristic in survival analyses of breast cancer, rather than pool the two conditions, since prognostic markers and effects are expected to differ. In this work we restrict our analysis to the 2,287 ER positive cases, the larger of the two strata. Follow up work is planned to analyse the ER negative cases. The SEARCH study is approved by the Cambridgeshire 4 Research Ethics Committee; all participants provided written informed consent.

All analyses modelled breast cancer specific mortality, with survival times left truncated at 10 years. This period was chosen since decisions relating to adjuvant therapy are often taken according to time horizon of ten years. 11% of women suffered breast cancer specific mortality during this follow-up, among whom the median survival time from baseline was five years. 44% of women whose survival times were censored have not yet been followed up for ten years — the median follow up time among these women is seven and a half years. Data was available for the following known prognostic risk factors: tumour size and grade, number of positive lymph nodes, HER2 status, use of chemotherapy and hormone therapy, and whether the patient was screen detected (suggesting the cancer was caught at an early stage though screening status is associated with improved outcome independent of stage[44]). These covariates were adjusted for in all analyses. Metastasis is clearly also important for breast cancer prognosis, however, since very few women in SEARCH had metastatic breast cancer at baseline (18/2,287), we excluded it from the models to avoid convergence issues. A sensitivity analysis including metastasis showed no change to the results presented in this paper.

Tumour markers

Expression of a particular protein will naturally vary between people, and at different locations in the body, including within tumours. In this experiment we sought to ascertain expression levels of 73 pre-selected proteins in tumour samples taken at diagnosis (i.e. baseline) using a technique known as immunohistochemistry. Tumour samples were stained with commercially available antibodies which produce a coloration, observable under a microscope, on contact with the protein of interest. Experts scored the stained tumour samples for the proportion of coloured cells in the biopsy (i.e. expressing the protein) on a four-point scale, and for the average intensity of that colour on a six-point scale. In total, intensity scores were taken for 51 proteins, and proportion scores for 45, and both were available for the three CSC markers. In addition, expression of various markers of immune infiltration, including CD8 and FOXP3, were measured in tumour associated lymphocytes. In situ hybridisation methods for detection of micro RNAs were implemented as previously described[45]. In total, 119 tumour markers were available for exploration for association with breast cancer survival. Correlations among the various tumour markers are shown in Figure 1a and 1b. Unsurprisingly, the two types of scoring (intensities and proportions of expression) are generally strongly correlated when measured for the same protein (Figure 1c). Unfortunately, there was substantial missingness among the tumour markers - see Figure 1d — with most missing for more than half the patients. There are two main reasons for missingness; by design and technical. Since the amount of biological material available for evaluation of novel tumour markers is limited, it was important to prioritise. Proteins were initially evaluated in a pilot study using only a subset of the available material. Based on preliminary analyses a judgment was taken whether to proceed to include all available material, hence in some instances only the data generated as part of the pilot study is available (the tumour markers with >70% missingness). The technical causes of missing data include biological variability e.g. differences in tumour size and dropout of samples during processing. This is a well-known unavoidable problem when tissue-microarrays (TMAs)[46] are used to evaluate large numbers of tumours. Fortunately, the correlation among the tumour markers (Figure 1a-c) enabled imputation of missing values - a description of how this was conducted, and how the multiply imputed datasets were analysed is given below in the methods.

3 Methods

3.1 The Weibull Regression Model

As noted above, we utilise the Weibull model in our sparse Bayesian regression framework for survival analysis. It is instructive to start with a description of the simpler exponential survival model, which the Weibull model extends. Under the exponential model, a patient i's hazard at time t is modelled as dependent on some P covariate values, denoted by vector $\boldsymbol{x}_i$, through an exponential link which ensures positivity of the hazard:

$$\lambda_i(t) = e^{\alpha + \boldsymbol{x}_i \boldsymbol{\beta}} = \lambda_i \qquad (1)$$

$\boldsymbol{\beta}$ in (1) is a P-length vector of covariate effects, and α denotes an intercept term. Note the lack of dependency on time in (1) — under the exponential model the hazard is assumed constant over time. The corresponding survival function, for example for patient i, is straight forward to derive as:

$$S(t) = e^{(-\int_0^t \lambda_i dx)} = e^{-t\lambda_i}$$

The assumption that hazard does not depend on time is likely to be overly simplistic for most real world scenarios. The Weibull model extends the exponential model by modifying the survival function with a parameter k as follows:

$$S(t) = e^{(-t\lambda_i)^k}$$

$k > 0$, known as the Weibull 'shape' parameter, induces a dependency between the hazard and time t:

$$\lambda(t) = -\frac{d}{dt} log(S(t)) = \lambda_i k (\lambda_i t)^{k-1}$$

Therefore if $k > 1$ the baseline hazard function increases as time progresses, but if $k < 1$ the hazard decreases.

Likelihood

Let vector $\boldsymbol{t}$ contain the observed survival times of n patients. Typically a study will not run long enough to observe whether or not the event occurs for each and every patient, resulting in so-called 'right censored' data. That is, for some patients we only know their minimum survival time. Therefore we also introduce an n-length vector of binary indicators $\boldsymbol{d}$ to capture, for each patient i, whether the event was observed during their follow up (in which case $d_i = 1$), or they were censored, in which case $d_i = 0$. If the event was observed for patient i (and $d_i = 1$), then t_i denotes their time to event. Otherwise, t_i denotes their length of follow up. The log-likelihood for parameters $\alpha, \boldsymbol{\beta}$ and k can be derived as:

$$log(L(\alpha, \boldsymbol{\beta}, k|\boldsymbol{t}, \boldsymbol{X})) = \sum_{i=1}^{n} d_i[\log(k) + k\log(\lambda_i) + (k-1)\log(t_i)] + (-t_i\lambda_i)^k \tag{2}$$

where λ_i is defined in (1).

3.2 Sparse Bayesian Weibull Regression (SBWR)

We present a full Reversible Jump MCMC algorithm for fitting Weibull survival models, in order to perform variable selection among the tumour markers. Henceforth we will refer to this framework, described below, as SBWR (Sparse Bayesian Weibull Regression).

We start by noting that baseline variables age, whether the patient was detected via a screening programme, chemotherapy treatment, hormone therapy, the number of positive lymph nodes and tumour size were excluded from the model selection framework and fixed to be included in the model at all times. Let vector $\boldsymbol{\delta}$ denote the log-hazard ratios associated with these 'fixed effects', and vector $\boldsymbol{z}_i$ denote the corresponding covariate values for patient i. Going forward, vector $\boldsymbol{x}_i$ will be used to denote patient i's tumour marker covariates only, and vector $\boldsymbol{\beta}$ the tumour marker log hazard ratios. P, the length of each of these vectors, therefore now denotes the number of tumour markers we wish to perform variable selection over. Under Reversible Jump, variable selection is facilitated by placing a prior density on $\boldsymbol{\beta}$ which depends on a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_P)$ of indices indicating whether each covariate is included in the model. For covariate p, $\gamma_p = 1$ indicates inclusion in the model and therefore that $\beta_p \neq 0$. Conditional on the latent variable $\boldsymbol{\gamma}$, i.e. a specific selection of tumour markers included in the model, patient i's hazard may now be written as:

$$\lambda_i|\boldsymbol{\gamma} = e^{\alpha+\boldsymbol{z}_i\boldsymbol{\delta}+\boldsymbol{x}_{i,\gamma}\boldsymbol{\beta}_\gamma}$$

where vector $\boldsymbol{\beta}_\gamma$ contains only the non-zero elements of $\boldsymbol{\beta}$, and vector $\boldsymbol{x}_{i,\gamma}$ contains patient i's corresponding subset of covariate values. The non-zero coefficients are assigned independent normal priors centred on 0, with a common variance σ^2_β:

$$p(\beta_p|\gamma_p = 1, \sigma_\beta) = N(0, \sigma^2_\beta) \text{ for } p = 1, \ldots, P \tag{3}$$

Rather than fixing σ_β, which controls the magnitude of included effects and therefore can have an important impact on the efficiency of the algorithm, we use a flexible hyperprior to allow adaption to the data at hand. We start by noting that all tumour marker covariates were normalised prior to analysis, so that (during modelling) all hazard ratios correspond to a standard deviation increase in the underlying variable. We chose a relatively informative Uniform(0,2) prior for σ_β. This has an expectation/median at 1, which would correspond to a prior with a 95% credible interval supporting hazard ratios between 0.14 and 7.12. However, this choice equally supports much smaller values of σ_β, (which would result in more pessimistic priors) as well as values up to the maximum of 2, which corresponds to a prior with a 95% credible interval supporting hazard ratios between 0.02 and 50.9 — well outside the range we realistically expect to observe. The 'fixed effects' $\boldsymbol{\delta}$ were ascribed weakly informative fixed $N(0, 10)$ priors rather than the hierarchical priors in (3). Since these covariates have well established associations with breast cancer survival they clearly do not have exchangeable effects a priori with the tumour markers, and so should not contribute to estimation of σ^2_β.

The model selection framework is completed by choosing a prior for $\boldsymbol{\gamma}$, the selection of tumour markers included in the model. We used a beta-binomial prior as described by Kohn et al.[47]:

$$p(\boldsymbol{\gamma}) = \int p(\boldsymbol{\gamma}|\omega)p(\omega)d\omega = \frac{B(p_\gamma + a_\omega, P - p_\gamma + b_\omega)}{B(a_\omega, b_\omega)} \tag{4}$$

where B is the Beta function and p_γ is the number of non-zero elements in $\boldsymbol{\gamma}$. Formally, $p_\gamma = \gamma^T I_P$ where I_P is the $P \times P$ identify matrix. Conceptually, ω denotes the underlying probability that each covariate has a non-zero effect, i.e. is included in $\boldsymbol{\gamma}$. Conditional on ω, all models of the same dimension are assumed, under this setup, equally likely a priori. a_ω and b_ω parameterise a Beta hyper-prior on ω. Since all tumour characteristics considered here were carefully selected for possible involvement in disease pathology, we set $a_\omega = 1$ and $b_\omega = 4$ which results in a prior on the probability of a true effect centred at 20%. Note, however, that this is only weakly informative due to the modest magnitudes of a_ω and b_ω relative to the number of tumour markers being analysed; ω should largely be learned from the data.

Finally, we must specify priors for the intercept α and the Weibull shape parameter k. In the spirit of Abrams et al., who provide a detailed discussion of fitting Weibull models in the Bayesian framework[48], we place normal priors with very large variance on α and on $log(k)$ (the log scale is used to ensure $k > 0$) which approximate 'reference' uniform priors over the entire real line;

$$p(\alpha) = N(0, 10^6)$$

$$p(log(k)) = N(0, 10^6)$$

3.3 Model fitting

As noted above, we cannot calculate the posterior of such a model analytically and so use Reversible Jump MCMC to sample from the required posterior[20]. The Reversible Jump sampling scheme starts at an initial model and corresponding set of parameter values, denote these $\boldsymbol{\gamma}(0)$ and $\boldsymbol{\theta}(0)$ respectively. To sample the next model and set of parameters, which we denote $\boldsymbol{\gamma}(1)$ and $\boldsymbol{\theta}(1)$, we propose moving from the current state to another model and/or set of parameter values, $\boldsymbol{\gamma}*$ and $\boldsymbol{\theta}^*$, by using a proposal function $q(\boldsymbol{\gamma}*, \boldsymbol{\theta}^*|\boldsymbol{\gamma}, \boldsymbol{\theta})$. We then accept these proposed values as the next sample with probability equal to the Metropolis-Hastings ratio:

$$MHR = \frac{P(D|\boldsymbol{\gamma}*, \boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*|\boldsymbol{\gamma}*)p(\boldsymbol{\gamma}*)}{P(D|\boldsymbol{\gamma}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})} \times \frac{q(\boldsymbol{\gamma}, \boldsymbol{\theta}|\boldsymbol{\gamma}*, \boldsymbol{\theta}^*)}{q(\boldsymbol{\gamma}*, \boldsymbol{\theta}^*|\boldsymbol{\gamma}, \boldsymbol{\theta})}$$

where D is the data, $P(D|.)$ is the Weibull likelihood function described in (2), $p(\boldsymbol{\theta}|\boldsymbol{\gamma})$ is the prior distribution of the parameters conditional on (that is, included in) the model, and $p(\boldsymbol{\gamma})$ is the model space prior defined in (4) . Therefore the proposed model and new parameter values are accepted with a probability proportional to their likelihood and prior. If this new set of values is accepted, the proposed set is accepted as $\boldsymbol{\gamma}(1)$ and $\boldsymbol{\theta}(1)$; otherwise, the sample value remains equal to the current sample value, i.e., $\boldsymbol{\gamma}(1) = \boldsymbol{\gamma}(0)$ and $\boldsymbol{\theta}(1) = \boldsymbol{\theta}(0)$. It can be shown that this produces a sequence of parameter samples that converge to the required posterior distribution[20]. The algorithm was implemented in Java; for technical details, for example the proposal distributions, we refer readers to the supplementary methods.

3.4 Post Processing

For all SBWR analyses of datasets with 119 covariates presented, i.e. the SEARCH data and the simulated datasets, the algorithm was run for 1 million iterations, after a burn-in of 1 million iterations, generating samples of all parameters. For the high-dimensional simulated datasets described below, the algorithm was run for 5 million iterations, after a burn-in of 5 million iterations for 10,000 covariates, and 10 million iterations after a burn-in of 10 million iterations for 20,000 covariates. These run lengths were deliberately longer than necessary for convergence, which was assessed using autocorrelation plots of the variable selections (see Supplementary Figure S2), chain plots of parameter values over the RJMCMC iterations (Supplementary Figure S3), and comparison of posterior probabilities obtained using different RJMCMC chains (Supplementary Figure S5). For each tumour marker covariate, complementary output was produced: The marginal posterior probability of inclusion, and the posterior median hazard ratio (and 95% credible interval) conditional on inclusion in the model. Furthermore, we obtain the posterior probability of any particular model, i.e. combination of covariates.

3.5 Multiple Imputation for Missingness

As noted above, there is substantial missingness among the covariates in the SEARCH breast cancer dataset. Since our algorithm currently cannot handle missingness, we proceeded to impute the missing values prior to analysis using multiple imputation by chained equations (MICE)[49, 50]; a well established and popular method of imputing missing data[51]. The MICE algorithm proceeds as follows. Initially, all missing values are filled in at random. Then, the first covariate with missing values, x_1 say, is regressed on all other covariates (and outcome), restricted to individuals with x_1 observed. The missing x_1 values are then updated with posterior predictive simulations from the resulting fitted model. This process is repeated for each covariate in turn to complete the first ‘cycle’. Subsequently, for each imputed dataset, 10 more ‘cycles’ were run to stabilise results. The entire procedure is then repeated independently M times resulting in a collection of completed datasets, the differences between which reflect uncertainty in the imputed values. We generated 20 imputed datasets in this manner using the STATA package ‘ICE’[52]. The choice of imputation models fitted for each covariate depend on the nature of its distribution. For the tumour markers, which are measured on an ordered categorical scale, ordinal regression was used to generate their posterior predictive distributions each 'cycle'. Likewise, ordinal regression was used for tumour size and grade, and positive lymph nodes which were treated as ordered categorical variables in this analysis. For the binary variables chemotherapy and screen detection logistic regression was used, and for morphology — an unordered categorical variable — multinomial logistic regression was used.

Bayesian analysis of multiply imputed datasets

To analyse multiply imputed datasets in a Bayesian framework, we follow the approach suggested by Gelman et al.[53] which is to (i) simulate many draws from the posterior distribution in each imputed dataset and (ii) mix all resulting draws into a single posterior sample. This final 'super' posterior therefore reflects the imputation uncertainty due to the heterogeneity among the chain-specific posteriors which have been pooled together.

3.6 Complementary Pairs Stability Selection

In the following sections, we will compare our method against a stability selection strategy utilising penalised regression of the LASSO form[1]. Stability selection was recently popularised by Meinshausen and Buhlmann[11] and aims to improve the selection of variables provided by penalised regression methods by adding a resampling step which involves repeating the variable selection procedure (in our case, LASSO regression) in a large number of datasets randomly sampled from the original. For each subsample analysis the covariates selected and rejected by LASSO are recorded. 'Selection probabilities' are then calculated across the results of all sub-sampled datasets. Intuitively this provides a measure of significance for each covariate since the strongest signals should be more robust to perturbations of the data. Theoretical results have been derived which offer upper bounds on the number of 'noise' variables for various thresholds on these selection probabilities, allowing inference of statistically significant predictors[11]. These results were recently improved upon by Shah and Samworth[12], who propose sub-sampling exactly half the data for each subset analysis and, each time this done, analysing both halves (i.e. the two complementary pairs) of the partitioned dataset. They provide a novel set of theoretical results to estimate the rate of ‘noise’ variables selected at different thresholds on the resulting selection probabilities. Their method leads to less conservative selections of covariates — a known issue with stability selection[12].

4 Simulation Study

In this section we used simulated data to investigate the performance of SBWR posterior probabilities in identifying true signal variables from noise variables. We compared performance against the selections provided by LASSO cox regression with the penalty parameter set to the optimum under ten fold cross-validation, and against the selection probabilities from Lasso Cox regression under CPSS.

4.1 Generation of the simulated data

Initially, simulated datasets were designed to have the same number of patients and covariates as the SEARCH breast cancer dataset, and the same real-life correlation structure as amongst the tumour markers. Hence, the covariate matrix of 119 tumour markers among the 2,287 ER positive patients was used from the SEARCH dataset in each replicate simulated dataset. We chose to ignore the missingness in the real data for the simulation study, simply to avoid the computational burden that would have arisen if multiple imputation chains were analysed for each replicate simulated dataset. Missing covariate values were filled in, arbitrarily, from the first multiple imputation chain.

Generation of simulated survival outcomes

We simulated outcome data according to the Generalized gamma parametric survival model, a flexible framework encompassing four of the commonly used parametric survival models (exponential, Weibull, log-normal and gamma) as special cases[54, 55]. In comparison to the Weibull, the Generalized gamma uses an extra parameter to model the hazard function, thus enabling a wider range of survival trajectories to be captured. Using the parameterisation of Prentice [54], in terms of three parameters μ, σ and q, the hazard function is described by:

$$\lambda(t) = \frac{|q|}{\sigma t\Gamma q^-2} exp(q^{-2}(qw - e^{qw}))(q^{-2})^{q^{-2}}$$

when $q \neq 0$ and where $w = (log(t) - \mu)/\sigma$. When $q = 0$ the hazard function becomes:

$$\lambda(t) = \frac{1}{\sqrt{2\pi}\sigma t} exp\{-\frac{1}{2\sigma^2}(log(t) - \mu)^2\}$$

When $q = 1$, the Generalised gamma reduces to the Weibull with $k = 1/\sigma$ and $\lambda = exp(-\mu)$. For a more detailed description of the Generalized gamma and relationship with other survival models we recommend referring to Cox et al[55] and Jackson et al[56]. To capture associations between predictors and outcome the parameter μ may be substituted for the standard linear predictor. Therefore, to induce associations between the covariates and outcome in the simulated data we drew survival times from a Generalised gamma distribution with

$$\boldsymbol{\mu} = \alpha + \boldsymbol{\beta}\boldsymbol{X}$$

where the covariate matrix $\boldsymbol{X}$ is that of the real data from the first imputation chain, and $\boldsymbol{\beta}$ is a vector of 119 tumour marker effects on survival. Note that the effects in a Generalized gamma model do not correspond to hazard ratios since hazards are no longer proportional under the more complex likelihood. Since, however, the Generalized gamma model has the accelerated failure-time property they do still correspond to differences in expected survival time. For the covariate effects, $\boldsymbol{\beta}$, 12 were randomly selected (approximately 10%) to have 'true', i.e. non-zero, effects. This random selection was only carried out once and used for all the simulation scenarios described below. We wished to use realistic effect magnitudes and so assigned these parameters the 12 largest coefficients from one-at-a-time Generalized gamma regressions of each tumour marker in the real data. That is, effect sizes observed in the real data were used, but arbitrarily re-assigned to different covariates. The absolute values of the 12 non-null elements of $\boldsymbol{\beta}$ ranged from 0.25 to 0.38 (note that all covariates were standardised to have unit variance). To determine realistic values for the remaining parameters α, σ and q we fitted a Generalized gamma regression model including an intercept term only (i.e. the 'null' model) in the real dataset. The resulting estimates of α, σ and q (3.28, 0.80, and -2.19 respectively) were used in the subsequent simulations. As noted above, the Generalized gamma is equivalent to the Weibull when $q = 1$. Since we are using $q = -2.19$, the simulated survival times are not Weibull distributed. This was done on purpose so that the simulation setup does not give an unfair advantage to SBWR, the only of the three methods to use the parametric Weibull likelihood, rather than the semi-parametric Cox likelihood.

Survival times were drawn from a Generalized gamma distribution according to the resulting linear predictor and parameters described above, and truncated at 10 years to mimic the actual data. Survival status was set to 'survived' where the survival time exceeded 10 years (before truncation), and 'died' otherwise. Using the same parameters, survival outcomes were re-drawn 20 times to create 20 replicate simulated datasets of 2,287 'patients' each. This process was repeated to generate additional simulated datasets in which the covariate effect sizes used for the simulations were halved to create a harder problem for the regression models, and again setting all covariate effects to zero to examine performance under the null. Generalized gamma simulation draws and regression model fitting was carried out using the excellent R package 'flexsurv', developed by Jackson et al[56].

High Dimensional Data Simulations

We also expanded the simulation setup to explore the performance of our method in much larger datasets, that is with more samples than covariates. To this end, we duplicated the covariate matrix described above, $\boldsymbol{X}$, multiple times column-wise (i.e. to add covariates). Each time a duplicate was added the rows were randomised such that none of the newly added covariates would be co-linear with their counterparts in the original, and first instance of, $\boldsymbol{X}$. This process was repeated until $P = 10,000$ and $P = 20,000$ covariates were present, resulting in two new high-dimensional covariate matrices with 2,287 'patients' each. To clarify, these consist of 119-covariate wide blocks within which the correlation structure is that among the SEARCH tumour markers. Covariates within each block, however, are independent of covariates in all other blocks.

To investigate performance in the 'needle in a haystack' setting, outcomes were drawn exactly as above, with the same effects at the same 12 tumour markers (arbitrarily using the first instance of $\boldsymbol{X}$). All other covariates were assigned null effects such that only 12/10,000 and 12/20,000 covariates had effects in the resulting high-dimensional simulated datasets. As above, outcomes were drawn 20 times, and the process repeated for 'full size', 'half size' and no effects.

4.2 Analysis of simulated datasets

All simulation analyses were carried out on Intel Xeon E5-2640 2.50GHz processors.

LASSO Penalised Cox Regression

Each simulated dataset was analysed using LASSO penalised Cox regression as implemented in the excellent R package 'glmnet'[57]. 10-fold cross-validation was used to choose the penalisation coefficient, and the selection of variables at the resulting optimum was recorded. These analyses took less than a minute per replicate for 119 covariates, around 19 minutes for 10,000 covariates and and around 28 minutes for 20,000 covariates.

LASSO Penalised Cox Regression with CPSS

Each simulated data set was also analysed using LASSO penalised Cox regression under Complementary Pairs Stability Selection (CPSS). 50 sets of complementary pairs were used as recommended by Shah and Samworth[12]. For each of the resulting 100 subdatasets, as above, the LASSO penalisation coefficient was optimised under 10-fold cross-validation. The resulting covariate selection probabilities were recorded. These took about 3 minutes per replicate for 119 covariates, around 18 hours for 10,000 covariates and around 25 hours for 20,000 covariates.

SBWR

For the simulations we assumed a complete lack of prior information and set $a = 1$ and $b = 1$ in the beta-binomial prior on model space. This corresponds to a naive, weakly informative, Uniform prior on the probability for a covariate to be truly causal. For the analysis of datasets with 119 covariates, 2 million RJMCMC iterations were run which took around 1 hour per replicate. For analyses of datasets with 10,000 covariates, 10 million iterations were run (about 9hrs per replicate), and for the datasets with 20,000 covariates, 20 million iterations were run (about 28 hours per replicate).

4.3 ROC Analysis for selection of true effects

We used ROC analysis to assess the ability of each approach to discriminate the 12 true signal variables from the noise variables in the simulated datasets. The resulting ROC curves for the different scenarios are shown in Figure 2 and the corresponding AUCs in Table 1. When the number of covariates was equivalent to the SEARCH dataset (i.e. 119) and 'full size' effects were simulated, the CPSS and SBWR selection probabilities demonstrated excellent, and equal, performance, and modest improvements over the simple LASSO. Both achieved average Areas Under the ROC Curve (AUC)s of 0.99 for discriminating the 12 true signal variables from the noise variables, compared to 0.92 for the LASSO selections. Under the harder 'half size' log-HR scenario, the posterior probabilities from SBWR demonstrated marginally better discrimination than the CPSS selection probabilities (average ROC AUC 0.97 vs 0.93). Again, both beat the LASSO selections - this time by a more substantial margin - which achieved an average ROC AUC of 0.86. Relative performance was similar in the high-dimensional simulated datasets of 10,000 and 20,000 covariates. SBWR and CPSS selection probabilities consistently outperformed the LASSO selections for discriminating the 12 true effects, with equal performance for the 'full size' effect scenario, and a marginal improvement in SBWR performance for the'half size' effect scenario. Under 'full size' simulated effects, the ability of SBWR and CPSS to discriminate the 12 signal variables from noise remained excellent up to 20,000 covariates, with average ROC AUCs >0.95, and LASSO also performed well (AUCs >0.92). When 'half size' effects were used the performance of all methods remained strong up to 10,000 covariates, but deteriorated by 20,000 covariates - SBWR and CPSS AUCs dropped to 0.82 and 0.78 respectively, while the mean LASSO AUC dropped to 0.72.

4.4 Performance under the null

Table 1 also includes median, and 2.5th to 97.5th percentile ranges, of the CPSS and SBWR selection probabilities, and mean selection probabilities from the LASSO analyses across covariates and simulation replicates under the null. There was no obvious cause for concern from any of the methods. When 119 covariates were analysed, SBWR demonstrated the smallest selection rates (mean 0.14 compared to 0.48 from Lasso under CPSS, and 0.60 from Lasso), though it should be kept in mind these selection probabilities from Lasso and CPSS do not have the same interpretation as posterior probabilities. Performance of all methods under the null was superior in the high dimensional data, with mean rates all under $1E-3$.

5 Tumour markers of breast cancer Survival in SEARCH

In this section we apply SBWR to explore a collection of tumour markers for association with breast cancer survival using data from 2,287 ER positive women collected as part of the SEARCH study. In the first instance, we restricted the analysis to the 75 tumour markers for which the majority of values were observed rather than imputed (i.e. those with missingness less than 50% — see Figure 1d). Analyses were also conducted using LASSO Cox regression with CPSS, and standard Weibull regressions including both each tumour marker one-at-a-time, a straight forward strategy that might typically be used here, and all tumour markers at once in a 'saturated' model.

To account for data missingness, 20 multiply imputed datasets were analysed independently using SBWR, and posterior results pooled, as described in the methods. Similarly, LASSO regressions under CPSS were performed in each multiply imputed dataset, and the resulting selection probabilities were averaged. For the one-at-a-time and 'saturated' Weibull regressions, results from each imputation chain were combined using Rubin's rules, as is standard practice[58]. Known predictors number of positive lymph nodes, tumour size and grade, detection by screening, chemotherapy, hormone therapy and morphology were fixed to be adjusted for the SBWR and LASSO models at all times, and adjusted for in the one-at-a-time regressions, in addition to age of diagnosis and study entry delay as possible confounders. In all frameworks, the tumour markers were analysed as ordinal continuous, assuming additive relationships with log-Hazards across all levels of the scales used in their measurement. Number of positive lymph nodes, tumour size, tumour grade and diagnosis age were also modelled as ordinal continuous variables using the levels derived by Wishart et al. to provide the best fitting additive relationship with log-hazards using an independent collection of 5,694 breast cancer patients[37] (see Table 3 for the categorisations).

Evidence of association for each tumour marker under SBWR and CPSS are shown in Figure 3a-b. Under SBWR, there was strong evidence of protective effects at PDCD4 (HR: 0.75 (0.62, 0.89), MPPI=84%) and the proportion score for PGR (HR: 0.86 (0.80, 0.93), MPPI=92%), and of a risk effect of AURKA (HR: 1.30 (1.11, 1.51), MPPI=68%). These three tumour markers were selected simultaneously in most of the top 20 models, providing strong evidence they represent independent effects on survival (Table 2). The Bayesian false discovery rate among these three tumour markers was estimated to be 19% which, while larger than we might have hoped, is to be expected since none of their posterior probabilities are decisive (please see the supplementary methods for a formal definition of the Bayesian false discovery rate). There was another 'band' of tumour markers for which there was suggestive evidence of association: the intensity score for GATA3, the proportion score for BCL2, and the proportion score for CD8 had similar posterior probabilities between 26% and 30%. However, the false discovery rate estimate increases to 45% when these tumour markers are included in the selection. Detailed results for these top six tumour markers under SBWR, in addition to the fixed effects and key model parameters, are presented in Table 3. The posterior distribution across the number of tumour markers included by SBWR had a large weight at 5 and above, the posterior probability of which was 63% (Supplementary Figure S1). This suggests that while the model may not be able to clearly discriminate among the more weakly associated tumour markers, there is more signal among these tumour markers than the top three associations, and that a future predictive model may benefit from leniency in which tumour markers are included. Interestingly, key prognostic factors tumour grade and HER2 had weaker effects upon inclusion of the tumour markers, suggesting that part of their association with survival may be through the tumour markers measured in this study (Supplementary Table S1). Variable selection auto-correlation plots (Supplementary Figure S2) and trace plots (Supplementary Figure S3) were consistent with convergence. Reassuringly, Supplementary Figure S4 shows that results were indistinguishable using more optimistic beta-binomial prior parameter choices of $a_\omega = 1$, $b_\omega = 3$ (centring the prior proportion of signals on 1/3) and a more pessimistic choice of $a_\omega = 1$, $b_\omega = 9$ (centring the prior proportion of signals on 0.1), as well as between different chains and starting values (Supplementary Figure S5). Furthermore, the results for the top tumour markers were consistent between inclusion or exclusion of imputed data (Supplementary Figure S6).

Encouragingly, the CPSS analysis ascribed the strongest selection probabilities to the same top three tumour markers as SBWR. Furthermore, the estimated percentage of 'noise' variables among these proteins was similar to the Bayesian false discovery rate estimate at 21%. There was somewhat less separation among the CPSS selection probabilities (Figure 3b) such that other markers, which were assigned weaker evidence under SBWR, achieved similar selection probabilities to the top three signals. In the one-at-a-time regressions (Supplementary Figure S7a), as under SBWR, there was strong evidence for PDCD4 and the proportion score for PGR with p-values for association that easily surpassed a multiplicity adjusted Bonferroni threshold of 6.7×10^{-4} ($p = 4.6 \times 10^{-}5$ and $1.1 \times 10^{-}6$ respectively - Figure 3a). However, the intensity score for PGR, which was ruled out under SBWR as confounded by its strong association with the proportion score, also reached significance (p=1.9×10^{-4}). AURKA, which obtained strong evidence of association under SBWR, was not significant falling short of the Bonferroni threshold (p=$1.05 \times 10^{-}3$). As in this application the number of predictors is smaller than the number of subjects, we also estimated a saturated Weibull model which included all 75 tumour markers simultaneously. In the saturated regression (Supplementary Figure S7b) the proportion score for PGR also reached significance ($p = 0.024$). The only other marker to reach significance was an intensity score for GATA3 ($p = 0.040$); we expect this is a spurious result arising from overfitting due to use of the saturated model. The fact that AURKA and PDCD4 both received p-values greater than 0.05 is likely reflective of the increase in power using sparse models under SBWR and LASSO.

Finally, we repeated the SBWR and CPSS analyses of SEARCH, extending to the complete set of 119 tumour markers, i.e. including tumour markers for which more than 50% of values were imputed. Inference was unchanged for the previously analysed 75 tumour markers, and there was no compelling evidence for any of the newly included tumour markers (Supplementary Figure S9).

6 Discussion

As large data-rich studies become common place in medical research, there is a growing need for regression tools which can facilitate variable selection over many predictors.

Attractive features of developing solutions in the Bayesian sparse regression framework include adequate reflection of uncertainty in the selection of covariates through inference of posterior probabilities for each predictor and possible model and, perhaps most importantly, that prior information can potentially be naturally incorporated through additional modelling of ω in the spirit of Quintana and Conti for the linear model[59]. We present, to our knowledge, the first implementation of a Bayesian variable selection algorithm for survival analysis under the Weibull model.

Over a range of realistic simulation scenarios our method generally demonstrated similar performance, and at times a marginal improvement in specificity, in comparison to an alternative frequentist strategy — penalised Cox regression with stability selection to generate measures of evidence for each covariate (specifically Complementary Pairs Stability selection[12]). Our simulation study also demonstrated that the current implementation of our method can cope with high-dimensional data up to 20,000 predictors, with computational times similar to the stability selection based approach (approximately one day for n=2,287 on an Intel Xeon E5-2640 2.50GHz processor).

Subsequently, we conducted a real data application in which 119 prospectively measured immunohistochemical tumour markers were explored for their association with survival among 2,287 ER positive breast cancer cases. Three proteins stood out with evidence of independent effects; PDCD4, PGR and AURKA. Discrimination, i.e. separation between the top signals and other tumour markers, was clearer when using SBWR in comparison to CPSS, consistently with the specificity improvements observed in some of the simulation scenarios. We also compared our results with those from a univariate strategy that might typically be used to analyse such data, highlighting some of the benefits of multivariate modelling.

Of the top three proteins, two are becoming increasingly recognised as powerful prognostic factors in ER-positive breast cancer. Indeed most schemes for clinical classification of subgroups of breast tumours based on molecular profiles include PGR[43, 60, 61] and, more recently, by using PGR expression at a higher threshold it has been proposed that it ought to be used to identify indolent ER-positive 'luminal A' tumours in the clinical setting[62]. Following numerous high-resolution molecular profiling studies over the past decade, tumour cell proliferation has been confirmed as the most powerful predictor of outcome in ER-positive tumours[63, 64]. There are potentially dozens of methods for measuring tumour cell proliferation including assaying different molecular markers of cell cycle. We have previously conducted a systematic comparison of the relative prognostic power of a panel of six proteins associated with cell-cycle including AURKA[65]. This study, based on the SEARCH dataset, identified AURKA as most strongly associated with outcome, outperforming the other investigated markers including marker-combinations[65]. Moreover, at the level of mRNA, AURKA has been identified as a prototypical marker of proliferation and selected for optimal classification of breast tumours into distinct molecular subgroups[66]. PDCD4 has not been investigated as a potential prognostic marker in breast cancer previously. However studies in lung[67] and salivary gland tumours[68] have shown an association with outcome. It is a well-known tumour suppressor and thought to inhibit the translation of proteins by interacting with eukaryotic translation factor 4A (eIF4A)[69]. The strong independent association between PDCD4 and outcome revealed by this analysis is a novel finding which highlights PDCD4 as a potentially useful clinical marker of outcome requiring further evaluation. Interpretation of these results should, however, be mindful of the estimated false discovery rate, 19%, suggesting that up to one of the three proteins is expected to be a false positive.

A further caveat to our real data application is that our treatment of missing data may be sub-optimal. Missing data was imputed using the well established technique of multiple imputation using chained equations[51], after which posteriors were pooled from individual analyses of 20 chains as suggested by Gelman et al.[53]. However, after a simulation study on the practical performance of this approach, Zhou and Reiter concluded that 100 or more chains should be used to achieve adequate coverage of the target posterior[70]. We did not do so here due to the computational time required to run our algorithm that many times, and since sensitivity analyses using less chains showed no substantive difference in estimates and inference. We also note that, while it was advisable for penalised regression, and our approach due to the prior framework, in general one should be very careful normalising predictors by their standard deviation[71]. In our case, however, none of the top three markers had extreme standard deviations prior to normalisation (ranging from 0.85 to 1.72), so our key results should not be meaningfully impacted by this issue.

Although our algorithm was technically challenging to develop, since both models and parameters are sampled during Reversible Jump MCMC, the framework used for variable selection is relatively simplistic. First, we specified independent priors for all covariate effects. Ideally, a multivariate normal prior would be used to reflect that correlated covariates are likely to have correlated effects. Zellner proposed the use of 'g-priors' in which a multivariate normal is used as a prior for the regression coefficients with a correlation structure informed by that of the covariate matrix[72]. In the context of linear regression, g-priors also preserve the ability to use conjugate results for coefficient effects and have been successfully implemented in the ESS sparse Bayesian regression framework[28, 29, 30]. It is worth noting, however, that the SSS algorithm also uses independent priors[27], and the use of independent priors in the work we present here did not prove problematic. Nevertheless, we intend to incorporate a g-prior option in the future. Second, the parametric assumptions imposed on the hazard function under the Weibull model might be too restrictive for some problems. Haneuse et al. have proposed a flexible Bayesian approach for capturing much more complex hazard functions, including to account for potentially time varying predictor effects[73]. Future work could also involve incorporating their ideas into our algorithm resulting in a considerably more flexible tool for survival analysis.

Although the runtimes of our algorithm when applied to high-dimensional datasets of 20,000 covariates were similar to those from a state-of-the-art implementation of LASSO, there is certainly room for improvement. An alternative strategy might be to avoid Reversible Jump altogether, and induce sparsity via independent double-exponential Laplacian priors on the Weibull covariate effects; a so-called ‘Bayesian LASSO’ model due to demonstrated similarity of results with the LASSO[74]. This would sacrifice the arguably more natural prior setup of the beta-binomial which, for example, allows direct specification of priors on the proportions of associated covariates. However, removing Reversible Jump from the MCMC algorithm could considerably improve efficiency. Another way we might improve the efficiency of our algorithm could be to employ Evolutionary Monte Carlo scheme which has proved effective for exploring parameter spaces consisting of hundreds of thousands of predictors[28, 30]. We plan to investigate both strategies in future work.

Our method of variable selection is relevant both to breast cancer research and clinical practice. Cancer research has been transformed by the introduction of high-throughput technologies which enable scientists to interrogate all expressed genes in a tumour and, more recently, the sequence of the entire cancer genome at single nucleotide resolution in a single experiment[75]. This has led to a proliferation of large datasets comprising hundreds to tens of thousands of molecular features. The emergence of such abundant data poses a strategic problem for the cancer biologist: How best can a shortlist of molecules of probable importance be distilled from such a multitude? One approach has been to use a combination of biological knowledge and statistical inference[76]. However, an alternative may be to use a legitimate end-point such as disease-specific-survival to infer which of a set of molecules influences the clinical behaviour of a tumour and is, therefore, likely to reflect its biological characteristics. Variable selection which accounts for the relative contribution of each of a large number of predictors represents a powerful method for identifying candidate molecules which warrant further biological investigation. Those molecules which are confirmed by such work to play a key role in tumour progression would represent lucrative targets for novel therapies.

Accurate risk prediction in breast cancer is important since many therapies have a modest effect on mortality and the absolute benefit of such toxic therapies is dependent on absolute risk of relapse or death[40]. Therefore even modest improvements in risk-prediction can influence treatment decisions. Current clinical methods heavily rely on conventional clinical parameters to estimate risk such as tumour size and grade, which are already measured rigorously and are not likely to be much improved[38]. However, molecular characteristics of tumours are not much utilised and represent an important avenue for improving our approach. The impact of abundant molecular data on clinical practice has been facilitated by studies over the past decade which used frequentist approaches to compile risk-prediction signatures for certain clinical endpoints[77]. These methods have had varying success and do not systematically account for the relative contribution of different variables. Through systematic consideration of multivariate models which account for the dependencies between covariates, our method of variable selection is likely to highlight not only molecules of biological importance but also to improve current risk-prediction methods. These benefits will extend to all common solid tumours in addition to breast cancer.

In summary, we present a new implementation of a Reversible Jump MCMC algorithm for Bayesian variable selection in survival analysis under the Weibull regression model. We demonstrate equal, or marginally superior, sensitivity and specificity to an alternative state-of-the-art approach over a range of realistic simulation scenarios with up to 20,000 covariates. Finally, resulting from a real data application in which our method demonstrated superior specificity over alternative approaches, we present evidence for three possible prognostic tumour markers of breast cancer survival. Despite the conceptual limitations listed above, in practice our software proved reliable, robust and efficient across the range of analyses presented here. Furthermore, our current implementation offers enormous flexibility for incorporation of prior information on effect magnitudes (individual priors can be specified for every covariate), and on relative probabilities of effect - the model space may be partitioned into as many components as required, each with an individual prior on the expected number of effects. This could, for example, be utilised to reflect that a subset of features lie in a known pathway for the disease being modelled. We have incorporated the algorithm, which was developed in java, into a freely available and easy to use R package called 'R2BGLiMS'. For download and installation instructions, please look under 'Other R packages' on our software page http://www.mrc-bsu.cam.ac.uk/software/, or, alternatively, direct download of 'R2BGLiMS' is available via github https://github.com/pjnewcombe/R2BGLiMS.

Acknowledgements

PJN and SR were funded by the Medical Research Council. PJN also acknowledges partial support from the NIHR Cambridge Biomedical Research Centre. The authors thank Richard Samworth and Rajen Shah for discussions on Complementary Pairs Stability Selection, Alexandra Lewin for discussions on False Discovery Rate estimation, and James Wason for valuable feedback on a draft of this manuscript. The authors also thank Bin Liu, Will Howatt, Sarah-Jane Dawson and John Le Quesne for their work in the collection and generation of the data analysed here. Finally, PJN thanks John Whittaker and Claudio Verzilli for supervision, discussion and feedback on the initial development of the Java RJMCMC algorithm during his PhD.

References

[1] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B*, 58(1):267–288, 1996.

[2] J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, 2001.

[3] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B*, 67(2):301–320, 2005.

[4] H. Zou. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.*, 101(476):1418–1429, 2006.

[5] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B*, 67(1):91–108, 2005.

[6] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–21, 2009.

[7] C. M. Vignal, A. T. Bansal, and D. J. Balding. Using penalised logistic regression to fine map HLA variants for rheumatoid arthritis. *Ann. Hum. Genet.*, 75(6):655–64, 2011.

[8] J. Peng, J. Zhu, and A. Bergamaschi. Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. *Ann. Appl. Stat.*, 4(1):53–77, 2010.

[9] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B*, 68:49–67, 2006.

[10] J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *Ann. Stat.*, 41(3):1111–1141, 2013.

[11] N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B*, 72(4):417–473, 2010.

[12] R. D. Shah and R. J. Samworth. Variable selection with error control: another look at stability selection. *J. R. Stat. Soc. Ser. B*, 75(1):55–80, 2013.

[13] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the Lasso. *Ann. Stat.*, 42(2):413–468, 2014.

[14] J. Taylor, R. Lockhart, R. J. Tibshirani, and R. Tibshirani. Post-selection adaptive inference for Least Angle Regression and the Lasso. *arXiv:1401.3889v2*, 2014.

[15] A. Chatterjee and S. N. Lahiri. Bootstrapping Lasso Estimators. *J. Am. Stat. Assoc.*, 106(494):608–625, 2011.

[16] E. I. George and R. E. McCulloch. Variable Selection Via Gibbs Sampling. *J. Am. Stat. Assoc.*, 88(423):881–889, 1993.

[17] P. J. Brown, M. Vannucci, and T. Fearn. Multivariate Bayesian variable selection and prediction. *J. R. Stat. Soc. Ser. B*, 60(3):627–641, 1998.

[18] L. Kuo and B. Mallick. Variable selection for regression models. *Sankhy Indian J. Stat. Ser. B*, 60:65–81, 1998.

[19] F. Hoti and M. J. Sillanpää. Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity (Edinb).*, 97(1):4–18, 2006.

[20] P. J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711, 1995.

[21] D. J. Lunn, J. C. Whittaker, and N. Best. A Bayesian Toolkit for Genetic Association Studies. *Genet. Epidemiol.*, 30(3):231–247, 2006.

[22] C. Verzilli, T. Shah, and J. Casas. Bayesian meta-analysis of genetic association studies with different sets of markers. *Am. J. Hum. Genet.*, 82:859–872, 2008.

[23] P. J. Newcombe, C. Verzilli, J. P. Casas, A. D. Hingorani, L. Smeeth, and J. C. Whittaker. Multilocus Bayesian meta-analysis of gene-disease associations. *Am. J. Hum. Genet.*, 84(5):567–80, 2009.

[24] P. J. Newcombe, B. H. Reck, J. Sun, G. T. Platek, C. Verzilli, A. K. Kader, S.-T. Kim, F.-C. Hsu, Z. Zhang, S. L. Zheng, V. E. Mooser, D. Lynn, C. F. Spraggs, J. C. Whittaker, R. S. Rittmaster, and J. Xu. A Comparison of Bayesian and Frequentist Approaches to Incorporating External Information for the Prediction of Prostate Cancer Risk. *Genet. Epidemiol.*, 36:71–83, 2012.

[25] L. C. McCarthy, P. J. Newcombe, J. C. Whittaker, J. I. Wurzelmann, M. A. Fries, N. R. Burnham, G. Cai, S. W. Stinnett, T. M. Trivedi, and C.-F. Xu. Predictive models of choroidal neovascularization and geographic atrophy incidence applied to clinical trial design. *Am. J. Ophthalmol.*, 154(3):568–578.e12, 2012.

[26] E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Stat. Sin.*, 7:339–373, 1997.

[27] C. Hans, A. Dobra, and M. West. Shotgun Stochastic Search for Large p Regression. *J. Am. Stat. Assoc.*, 102(478):507–516, 2007.

[28] L. Bottolo and S. Richardson. Evolutionary Stochastic Search for Bayesian Model Exploration. *Bayesian Anal.*, 5(3):583–618, 2010.

[29] L. Bottolo, M. Chadeau-hyam, D. I. Hastie, S. R. Langley, E. Petretto, L. Tiret, D. Tregouet, and S. Richardson. ESS ++ : a C ++ objected-oriented algorithm for Bayesian stochastic search model exploration. *Bioinformatics*, pages 2–3, 2011.

[30] L. Bottolo, M. Chadeau-Hyam, D. I. Hastie, T. Zeller, B. Liquet, P. Newcombe, L. Yengo, P. S. Wild, A. Schillert, A. Ziegler, S. F. Nielsen, A. S. Butterworth, W. K. Ho, R. Castagné, T. Munzel, D. Tregouet, M. Falchi, F. Cambien, B. r. G. Nordestgaard, F. Fumeron, A. Tybjæ rg Hansen, P. Froguel, J. Danesh, E. Petretto, S. Blankenberg, L. Tiret, and S. Richardson. GUESS-ing Polygenic Associations with Multiple Phenotypes Using a GPU-Based Evolutionary Stochastic Search Algorithm. *PLoS Genet.*, 9(8):e1003657, 2013.

[31] E. Petretto, L. Bottolo, S. R. Langley, M. Heinig, C. McDermott-Roe, R. Sarwar, M. Pravenec, N. Hübner, T. J. Aitman, S. A. Cook, and S. Richardson. New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Comput. Biol.*, 6(4):e1000737, 2010.

[32] R. B. O'Hara and M. J. Sillanpaa. A Review of Bayesian Variable Selection Methods: What, How and Which. *Bayesian Anal.*, 4(1):85–118, 2009.

[33] N. Keiding, P. K. Andersen, and J. P. Klein. The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates. *Stat. Med.*, 16(1-3):215–24, 1997.

[34] D. Collett. *Modelling Survival Data in Medical Research, Second Edition.* CRC Press, 2003.

[35] R. Peto, J. Boreham, M. Clarke, C. Davies, and V. Beral. UK and USA breast cancer deaths down 25% in year 2000 at ages 2069 years. *Lancet*, 355:1822, 2000.

[36] S. Mayor. UK deaths from breast cancer fall to lowest figure for 40 years. *BMJ*, 338:b1710, 2009.

[37] G. C. Wishart, E. M. Azzato, D. C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, and P. D. P. Pharoah. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Res.*, 12(1):R1, 2010.

[38] G. C. Wishart, C. D. Bajdik, E. Dicks, E. Provenzano, M. K. Schmidt, M. Sherman, D. C. Greenberg, A. R. Green, K. A. Gelmon, V.-M. Kosma, J. E. Olson, M. W. Beckmann, R. Winqvist, S. S. Cross, G. Severi, D. Huntsman, K. Pylkäs, I. Ellis, T. O. Nielsen, G. Giles, C. Blomqvist, P. A. Fasching, F. J. Couch, E. Rakha, W. D. Foulkes, F. M. Blows, L. R. Bégin, L. J. van't Veer, M. Southey, H. Nevanlinna, A. Mannermaa, A. Cox, M. Cheang, L. Baglietto, C. Caldas, M. Garcia-Closas, and P. D. P. Pharoah. PREDICT Plus: development and validation of a prognostic model for early breast cancer that includes HER2. *Br. J. Cancer*, 107(5):800–7, 2012.

[39] C. Curtis, S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, S. Gräf, G. Ha, G. Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerø d, A. Green, E. Provenzano,

G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A.-L. Bø rresen Dale, J. D. Brenton, S. Tavaré, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–52, 2012.

[40] R. Peto, C. Davies, J. Godwin, R. Gray, H. C. Pan, M. Clarke, D. Cutter, S. Darby, P. McGale, C. Taylor, Y. C. Wang, J. Bergh, A. Di Leo, K. Albain, S. Swain, M. Piccart, and K. Pritchard. Comparisons between different polychemotherapy regimens for early breast cancer: meta-analyses of long-term outcome among 100,000 women in 123 randomised trials. *Lancet*, 379(9814):432–44, 2012.

[41] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–6, 2002.

[42] M. Zöller. CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? *Nat. Rev. Cancer*, 11(4):254–67, 2011.

[43] F. M. Blows, K. E. Driver, M. K. Schmidt, A. Broeks, F. E. van Leeuwen, J. Wesseling, M. C. Cheang, K. Gelmon, T. O. Nielsen, C. Blomqvist, P. Heikkilä, T. Heikkinen, H. Nevanlinna, L. A. Akslen, L. R. Bégin, W. D. Foulkes, F. J. Couch, X. Wang, V. Cafourek, J. E. Olson, L. Baglietto, G. G. Giles, G. Severi, C. A. McLean, M. C. Southey, E. Rakha, A. R. Green, I. O. Ellis, M. E. Sherman, J. Lissowska, W. F. Anderson, A. Cox, S. S. Cross, M. W. R. Reed, E. Provenzano, S.-J. Dawson, A. M. Dunning, M. Humphreys, D. F. Easton, M. García-Closas, C. Caldas, P. D. Pharoah, and D. Huntsman. Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. *PLoS Med.*, 7(5):e1000279, 2010.

[44] S.-J. Dawson, N. Makretsov, F. M. Blows, K. E. Driver, E. Provenzano, J. Le Quesne, L. Baglietto, G. Severi, G. G. Giles, C. A. McLean, G. Callagy, A. R. Green, I. Ellis, K. Gelmon, G. Turashvili, S. Leung, S. Aparicio, D. Huntsman, C. Caldas, and P. Pharoah. BCL2 in breast cancer: a favourable prognostic marker across molecular subtypes and independent of adjuvant therapy received. *Br. J. Cancer*, 103(5):668–75, 2010.

[45] J. L. Quesne, J. Jones, J. Warren, S.-J. Dawson, H. R. Ali, H. Bardwell, F. Blows, P. Pharoah, and C. Caldas. Biological and prognostic associations of miR-205 and let-7b in breast cancer revealed by in situ hybridization analysis of micro-RNA expression in arrays of archival tumour tissue. *J. Pathol.*, 227(3):306–14, 2012.

[46] J. Kononen, L. Bubendorf, A. Kallioniemi, M. Bärlund, P. Schraml, S. Leighton, J. Torhorst, M. J. Mihatsch, G. Sauter, and O. P. Kallioniemi. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.*, 4(7):844–7, 1998.

[47] R. Kohn, M. Smith, and D. Chan. Nonparametric regression using linear combinations of basis functions. *Stat. Comput.*, 11:313–322, 2001.

[48] K. Abrams, D. Ashby, and D. Errington. A Bayesian approach to Weibull survival models–application to a cancer clinical trial. *Lifetime Data Anal.*, 2(2):159–74, 1996.

[49] J. L. Schafer. Multiple imputation: a primer. *Stat. Methods Med. Res.*, 8:3–15, 1999.

[50] S. van Buuren, H. C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Stat. Med.*, 18(6):681–94, 1999.

[51] J. A. C. Sterne, I. R. White, and J. B. Carlin. Multiple imputation for missing data in epidemiological and clinical research : potential and pitfalls. *BMJ*, 339(July):157–160, 2009.

[52] P. Royston and I. White. Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *J. Stat. Softw.*, 45(4), 2011.

[53] A. Gelman. *Bayesian Data Analysis.* Chapman & Hall/CRC, 2nd edition, 2004.

[54] R. Prentice. A log gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):539–544, 1974.

[55] C. Cox, H. Chu, M. F. Schneider, and A. Muñoz. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Stat. Med.*, 26(23):4352–4374, 2007.

[56] C. H. Jackson, L. D. Sharples, and S. G. Thompson. Survival Models in Health Economic Evaluations: Balancing Fit and Parsimony to Improve Prediction. *Int. J. Biostat.*, 6(1), 2010.

[57] N. Simon and J. Friedman. Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent. *J. Stat. Softw.*, 39(5):1–13, 2011.

[58] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley and Sons, New York, 1987.

[59] M. a. Quintana and D. V. Conti. Integrative variable selection via Bayesian model uncertainty. *Stat. Med.*, 32(28):4938–53, 2013.

[60] T. O. Nielsen, F. D. Hsu, K. Jensen, M. Cheang, G. Karaca, Z. Hu, T. Hernandez-Boussard, C. Livasy, D. Cowan, L. Dressler, L. A. Akslen, J. Ragaz, A. M. Gown, C. B. Gilks, M. van de Rijn, and C. M. Perou. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin. Cancer Res.*, 10(16):5367–74, 2004.

[61] L. A. Carey, C. M. Perou, C. A. Livasy, L. G. Dressler, D. Cowan, K. Conway, G. Karaca, M. A. Troester, C. K. Tse, S. Edmiston, S. L. Deming, J. Geradts,

M. C. U. Cheang, T. O. Nielsen, P. G. Moorman, H. S. Earp, and R. C. Millikan. Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *J. Am. Med. Assoc.*, 295(21):2492–502, 2006.

[62] A. Prat, M. C. U. Cheang, M. Martín, J. S. Parker, E. Carrasco, R. Caballero, S. Tyldesley, K. Gelmon, P. S. Bernard, T. O. Nielsen, and C. M. Perou. Prognostic significance of progesterone receptor-positive tumor cells within immunohistochemically defined luminal A breast cancer. *J. Clin. Oncol.*, 31(2):203–9, 2013.

[63] P. Wirapati, C. Sotiriou, S. Kunkel, P. Farmer, S. Pradervand, B. Haibe-Kains, C. Desmedt, M. Ignatiadis, T. Sengstag, F. Schütz, D. R. Goldstein, M. Piccart, and M. Delorenzi. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.*, 10(4):R65, 2008.

[64] C. Desmedt, B. Haibe-Kains, P. Wirapati, M. Buyse, D. Larsimont, G. Bontempi, M. Delorenzi, M. Piccart, and C. Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.*, 14(16):5158–65, 2008.

[65] H. R. Ali, S.-J. Dawson, F. M. Blows, E. Provenzano, P. D. Pharoah, and C. Caldas. Aurora kinase A outperforms Ki67 as a prognostic marker in ER-positive breast cancer. *Br. J. Cancer*, 106(11):1798–806, 2012.

[66] B. Haibe-Kains, C. Desmedt, S. Loi, A. C. Culhane, G. Bontempi, J. Quackenbush, and C. Sotiriou. A three-gene model to robustly identify breast cancer molecular subtypes. *JNCI*, 104(4):311–25, 2012.

[67] Y. Chen, T. Knösel, G. Kristiansen, A. Pietas, M. E. Garber, S. Matsuhashi, I. Ozaki, and I. Petersen. Loss of PDCD4 expression in human lung cancer correlates with tumour progression and prognosis. *J. Pathol.*, 200(5):640–6, 2003.

[68] C. Qi, Y. Shao, N. Li, C. Zhang, M. Zhao, and F. Gao. Prognostic significance of PDCD4 expression in human salivary adenoid cystic carcinoma. *Med. Oncol.*, 30(1):491, 2013.

[69] B. Lankat-Buttgereit and R. Göke. The tumour suppressor Pdcd4: recent advances in the elucidation of function and regulation. *Biol. Cell*, 101(6):309–17, 2009.

[70] X. Zhou and J. P. Reiter. A Note on Bayesian Inference After Multiple Imputation. *Am. Stat.*, 64(2):159–163, 2010.

[71] J. Bring. How to Standardize Regression Coefficients. *Am. Stat.*, 48(3):209–213, 1994.

[72] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Geol and A. Zellner, editors, *ayesian Inference Decis. Tech. Essays Honour Bruno Finetti*, pages 233–243. Elsevier Science, North Holland, Amsterdam, 1986.

[73] S. J.-P. A. Haneuse, K. D. Rudser, and D. L. Gillen. The separation of timescales in Bayesian survival modeling of the time-varying effect of a time-dependent exposure. *Biostatistics*, 9(3):400–10, 2008.

[74] M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 2008.

[75] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler. Cancer genome landscapes. *Science (80-.).*, 339(6127):1546–58, 2013.

[76] M. S. Lawrence, P. Stojanov, C. H. Mermel, J. T. Robinson, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 2014.

[77] J. S. Reis-Filho and L. Pusztai. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet*, 378(9805):1812–23, 2011.

7 Tables

Table 1: Comparison of methods in simulated data. The top part of the table presents areas under the receiver operator characteristic curve (ROC AUCs) for detection of the 12 true effects among the variables analysed. Results are averaged over the analysis of 20 replicate datasets for each simulation scenario, with the standard deviation across replicates included in brackets. The bottom part of the table presents mean selection rates of each method under the null, over all covariates and all simulation replicates, with the standard deviation included in brackets.

	LASSO	CPSS	SBWR
ROC AUCs			
119 covariates, 'full size' effects	0.92 (0.04)	0.99 (0.01)	0.99 (0.01)
119 covariates, 'half size' effects	0.86 (0.06)	0.93 (0.05)	0.97 (0.03)
10,000 covariates, 'full size' effects	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)
10,000 covariates, 'half size' effects	0.95 (0.05)	0.95 (0.03)	0.99 (0.01)
20,000 covariates, 'full size' effects	0.92 (0.02)	0.95 (0.02)	0.96 (0.04)
20,000 covariates, 'half size' effects	0.72 (0.05)	0.78 (0.03)	0.82 (0.15)
Selection rates under the null			
119 covariates	0.60 (0.49)	0.48 (0.32)	0.14 (0.25)
10,000 covariates	9.2E-4 (0.03)	6.9E-4 (5.0E-3)	1.6E-6 (2.1E-5)
20,000 covariates	0 (0)	1.8E-4 (1.8E-3)	0 (0)

Table 2: Top 20 models from the SBWR analysis of the SEARCH dataset, inferred by SBWR.

AURKA_P	BCL2_P	CK56	GATA3_I	PDCD4_{O2}	PGR_P	SMAD2_I	PTEN_{I3}	PTEN_{P3}	SLC7A5_P	CD8_P	Posterior Probability
●				●	●						4.5%
				●	●						2.5%
●				●	●					●	2.0%
●	●			●	●						1.3%
●			●	●	●						1.2%
				●	●					●	1.2%
●				●	●		●				0.7%
					●						0.6%
●			●		●						0.6%
●				●	●				●		0.5%
	●			●	●						0.5%
●	●			●							0.5%
●	●			●	●					●	0.5%
			●		●						0.4%
●					●						0.4%
●				●	●			●			0.4%
			●	●	●						0.4%
				●	●	●					0.3%
●				●	●	●					0.3%
●		●		●	●						0.3%

Table 3: SBWR results, for the fixed effects and top tumour markers associated breast cancer survival in SEARCH. †Modelled as ordinal continuous. ‡For the tumour markers, these were calculated conditional on inclusion in the model. *Marginal Posterior Probabilitiy of Inclusion in the model - may be interpreted as the posterior probability an association exists with survival, adjusted for all other covariates in the model.

	HR	95% CrI‡	MPPI*	Imputed
Fixed parameters				
Intercept	-7.32	(-8.07, -6.63)		0%
log(beta) Hyperprior SD (σ_β)	0.24	(0.12, 0.71)		0%
Weibull scale	1.74	(1.54, 1.96)		0%
Number Positive Nodes† (0, 1, 2-4, 5-9, 10+)	1.61	(1.45, 1.79)		8.4%
Tumour Size, mm† (<10, 10-19, 20-29, 30-49, 50+)	1.26	(1.09, 1.45)		3.8%
Tumour grade† (Low, Intermediate, High)	1.47	(1.14, 1.89)		10.5%
Morphology - Ductular	-	-		0%
Morphology - Lobular	1.55	(1.10, 2.16)		-
Morphology - Other	1.06	(0.64, 1.68)		-
HER2	1.47	(0.97, 2.18)		10.8%
Detection by screening	0.79	(0.55, 1.11)		6.1%
Hormone therapy	2.20	(1.38, 3.86)		<0.01%
Study entry delay, years	0.88	(0.79, 0.98)		0%
'Top' tumour markers				
PGR_P	0.86	(0.80, 0.93)	0.92	5.2%
PDCD4_{O2}	0.75	(0.62, 0.89)	0.84	43.2%
AURKA_P	1.30	(1.11, 1.51)	0.68	31.0%
CD8_P	0.92	(0.85, 0.98)	0.30	32.9%
GATA3_I	0.80	(0.68, 0.94)	0.30	41.6%
BCL2_P	0.94	(0.90, 0.98)	0.26	9.0%

8 Figures

Figure 1: Tumour marker correlation structure and missingness. a) A heatmap representing pairwise Pearson correlation statistics among the various IHC intensity score tumour markers. b) A heatmap representing pairwise Pearson correlation statistics among the various IHC proportion score tumour markers. c) Pearson correlation statistics between IHC intensity and proportion scores for those tumour markers where both were measured. d) Proportion of missing values for each tumour marker in the analysis population.

Figure 2: ROC analysis of simulation results. Performance of association measures from CPSS and SBWR, and the LASSO selections when the penalty parameter is set to the optimum under cross-validation, in distinguishing 12 signals from noise in datasets ranging from 119 to 20,000 covariates. Panels a, c and e show results from datasets simulated to have 12 'full size' effects, and panels b, d and f under 12 'half size' effects. Each ROC curve is vertically averaged over the results from the analysis of 20 replicate datasets.

Figure 3: Association of tumour markers with breast cancer survival in the SEARCH dataset. In all panels, each tumour marker is coloured according to its strongest pairwise correlation with one of the three top hits — PDCD4, PGR and AURKA. a) Selection probabilities from LASSO Cox regression with CPSS. b) SBWR posterior probabilities.

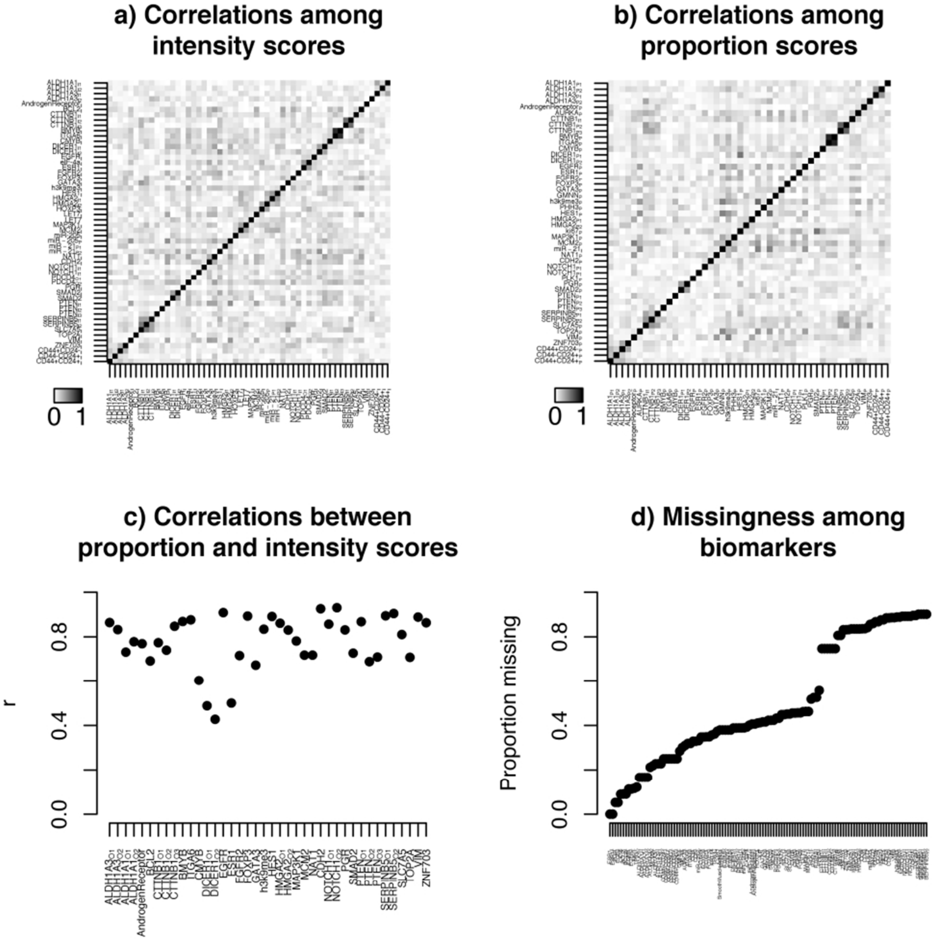

Figure 1: Tumour marker correlation structure and missingness. a) A heatmap representing pairwise Pearson correlation statistics among the various IHC intensity score tumour markers. b) A heatmap representing pairwise Pearson correlation statistics among the various IHC proportion score tumour markers. c) Pearson correlation statistics between IHC intensity and proportion scores for those tumour markers where both were measured. d) Proportion of missing values for each tumour marker in the analysis population.

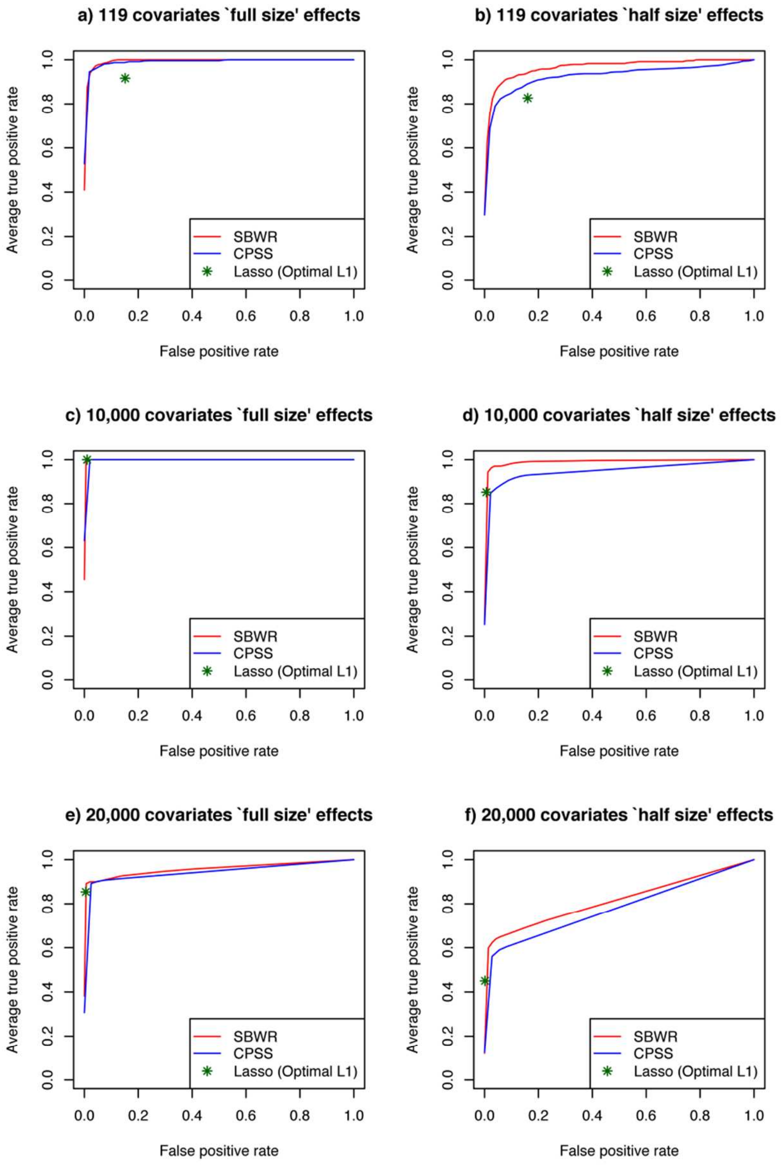

Figure 2: ROC analysis of simulation results. Performance of association measures from CPSS and SBWR, and the LASSO selections when the penalty parameter is set to the optimum under cross-validation, in distinguishing 12 signals from noise in datasets ranging from 119 to 20,000 covariates. Panels a), c) and e) show results from datasets simulated to have 12 ‘full size’ effects, and panels b), d) and f) under 12 ‘half size’ effects. Each ROC curve is vertically averaged over the results from the analysis of 20 replicate datasets.

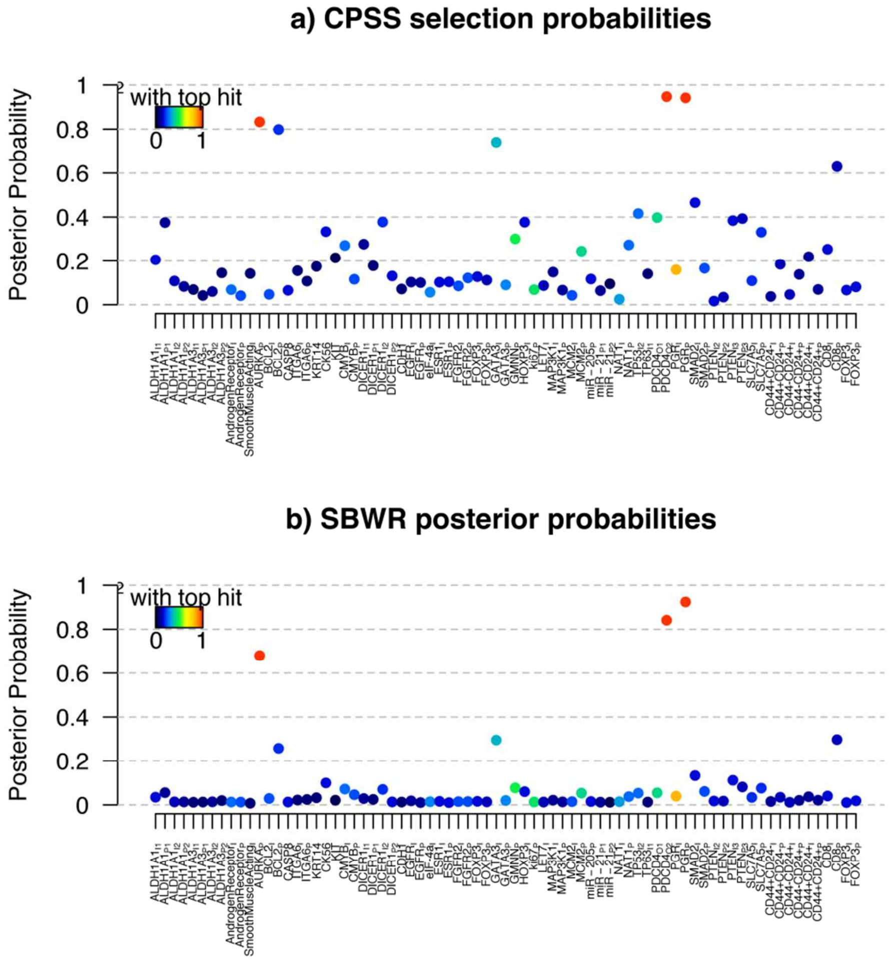

Figure 3: Association of tumour markers with breast cancer survival in the SEARCH dataset. In all panels, each tumour marker is coloured according to its strongest pairwise correlation with one of the three top hits - PDCD4, PGR and AURKA. a) Selection probabilities from LASSO Cox regression with CPSS. b) SBWR posterior probabilities.

Weibull regression with Bayesian variable selection to identify prognostic tumour markers of breast cancer survival: Supplementary Material

P. J . Newcombe[1], H. Raza Ali[2,3,4], F. M. Blows[5], E. Provenzano[6], P. D. Pharoah[4,5,7], C. Caldas[2,4,5], and S. Richardson[1]

[1]MRC Biostatistics Unit, Cambridge, UK
[2]Cancer Research UK Cambridge Institute
[3]Department of Pathology, University of Cambridge
[4]Cambridge Experimental Cancer Medicine Centre and NIHR Cambridge Biomedical Research Centre
[5]Department of Oncology, University of Cambridge
[6]NIH Cambridge Biomedical Research Centre
[7]Strangeways Research Laboratory

1 Supplementary Methods

1.1 Reversible Jump moves within the model space

The set of models that the Reversible Jump algorithm was allowed to mix between was defined by all possible combinations of tumour marker parameters being included or excluded for each marker. Therefore, when N_{Bio} are included in the analysis, there is a set of $\sum_{m=0}^{m=N_{Bio}} \binom{N_{Bio}}{m}$ possible models that the Reversible Jump algorithm may mix between.

Determining the probability of a model move was a two stage process. First, the type of move was determined from four possibilities: Adding a tumour marker HR, removing a tumour marker HR, swapping the HR of one tumour marker for another, or a 'null' move where no change is made. An addition can only occur when there are $< N_{Bio}$ HRs present, a removal can only occur when there are >0 HRs present, and a swap can only occur when there are >1 HRs present. Swap, addition and removal moves were given a $\frac{1}{6}$ probability each of happening, when available. The null move therefore had a $\frac{1}{2}$ probability of happening when all other move types were available, although this was increased by the probabilities assigned to other move types when they were unavailable. Second, if an addition, removal or swap move was selected, the tumour markers to be involved in the move were picked from the tumour markers available for the move (an addition, for example, can only choose from tumour markers not currently included) with equal probability.

In summary, the probability of a particular move within the model space is determined by multiplying the probability of the move type and, with the exception of a 'null' move, the probability of selecting the particular tumour marker(s) involved in the move.

1.2 Parameter Updates

We adopt a proposal mechanism that updates one parameter type each iteration of the Reversible Jump algorithm. For each proposal made in the reversible jump algorithm there are four types of parameters that may be updated:

- The intercept intercept α
- Tumour marker log-HRs $\boldsymbol{\beta}$.
- Tumour marker prior precision hyperparameter σ_β
- Weibull shape parameter k

The parameter type to update is chosen at random, with weighting equal to the number of occurrences of the parameter type in the model under consideration.

Updating the intercept α and tumour marker HRs $\boldsymbol{\beta}$

These are updated using a normal distribution centred on the current value (which may be 0 in the case of an 'addition' move). As explained above, only one log-HR is ever updated per iteration with the exception of a 'swap' move in which one log-HR is set to 0 while another has a new non-zero value proposed.

Updating the prior hyper parameter σ_{beta} and the Weibull shape parameter k

Since these parameters must always be positive, they are updated on the log-scale. This is also performed using a normal distribution centred on the current (log) value.

Adaption of proposal distribution variances

Each of the four proposal distributions described above is assigned a specific variance parameter. Small values result in a high number of proposals being accepted, since the values are closer to the previously accepted current values. However, another consequence is slower exploration of the posterior space such that the algorithm must be run for longer to obtain a representative sample. Therefore, in choosing the proposal variances it is important to find a balance between a reasonable acceptance rate of proposals and efficiency in traversing the posterior space. Each of the four variances are tuned during an initial 100,000 iteration adaption stage to target an acceptance rate of approximately 0.4 for each parameter type.

1.3 False Discovery Rates

To determine the statistical significance of the posterior probabilities of association for a covariate g, p_g, we calculated the 'Bayesian FDR' — first defined by Newton et al.[1], and further developed by Muller et al.[2] — for which there are several examples of successful application to real datasets[3, 4, 5]. Briefly, for a threshold p_{cut} on the posterior probabilities of association p_g, an empirical estimate of the FDR is given by:

$$1 - \frac{\sum_{g|p_g \geq p_{cut}} p_g}{n_{p_{cut}}}$$

where $n_{p_{cut}}$ is the number of variables with posterior probabilities above p_{cut}. Therefore, this is simply one minus the average posterior probability among variables which reach the threshold.

References

[1] M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–76, 2004.

[2] P. Müller, G. Parmigiani, and K. Rice. FDR and Bayesian multiple comparisons rules. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, Heckerman, A. D. Smith, and M. West, editors, *Bayesian Stat. Vol. 8*, pages 349–368. Oxford University Press, Oxford University Press, UK, 2007.

[3] P. Broët, A. Lewin, S. Richardson, C. Dalmasso, and H. Magdelenat. A mixture model-based strategy for selecting sets of genes in multiclass response microarray experiments. *Bioinformatics*, 20(16):2562–71, 2004.

[4] A. Lewin, S. Richardson, C. Marshall, A. Glazier, and T. Aitman. Bayesian modeling of differential gene expression. *Biometrics*, 62(1):1–9, 2006.

[5] A. Lewin, N. Bochkina, and S. Richardson. Fully Bayesian mixture model for differential gene expression: simulations and model checks. *Stat. Appl. Genet. Mol. Biol.*, 6(1):Article36, 2007.

2 Supplementary Tables

Table S1: Comparison of effect estimates for the fixed effects with and without tumour marker adjustment. The left two columns show results from running our SBWR algorithm on the SEARCH dataset with all covariates in the table fixed to be included, but the tumour markers excluded. The right two columns show results from the main SBWR analysis of SEARCH, in which the covariates in this table were fixed to be included, while simultaneously performing model selection over 75 tumour markers. Posterior results were pooled from individual analyses of 20 multiply imputed datasets, as described in the methods. †Modelled as ordinal continuous.

	Without tumour markers		With tumour markers	
	HR	95% CrI	HR	95% CrI
Number Positive Nodes† (0, 1, 2-4, 5-9, 10+)	1.62	(1.44, 1.82)	1.60	(1.42, 1.80)
Tumour Size, mm† (<10, 10-19, 20-29, 30-49, 50+)	1.21	(1.05, 1.39)	1.22	(1.06, 1.40)
Tumour grade† (Low, Intermediate, High)	1.52	(1.30, 1.77)	1.32	(1.11, 1.58)
Metastasis	4.06	(1.94, 7.69)	3.39	(1.55, 7.04)
Morphology - Ductular	-	-	-	-
Morphology - Lobular	1.37	(0.98, 1.88)	1.49	(1.05, 2.08)
Morphology - Other	0.93	(0.56, 1.49)	1.02	(0.61, 1.62)
HER2	1.61	(1.07, 2.32)	1.42	(0.93, 2.11)
Detection by screening	0.84	(0.59, 1.17)	0.80	(0.56, 1.14)
Hormone therapy	2.05	(1.29, 3.52)	2.17	(1.33, 3.80)
Study entry delay, years	0.87	(0.75, 0.99)	0.86	(0.74, 0.98)

3 Supplementary Figures

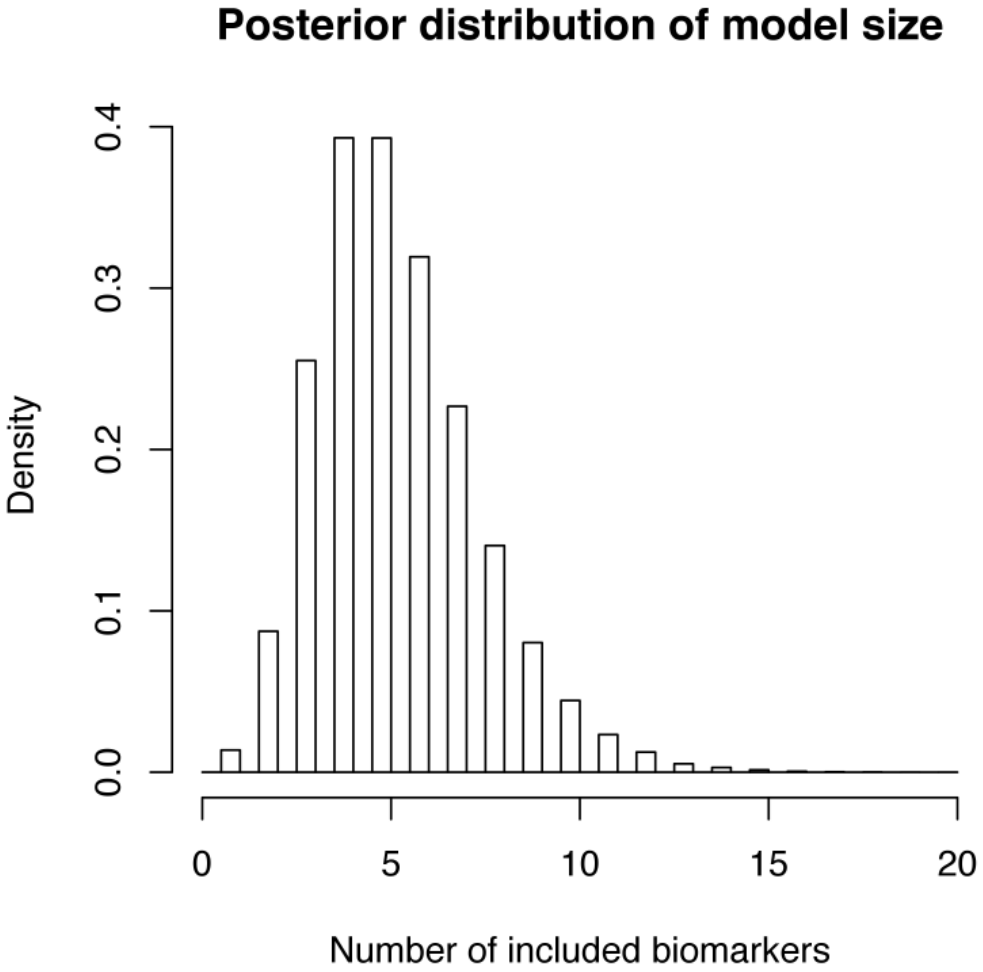

Figure S1: Posterior distribution of model size, i.e. number of included tumour markers, from the SBWR analysis of SEARCH.

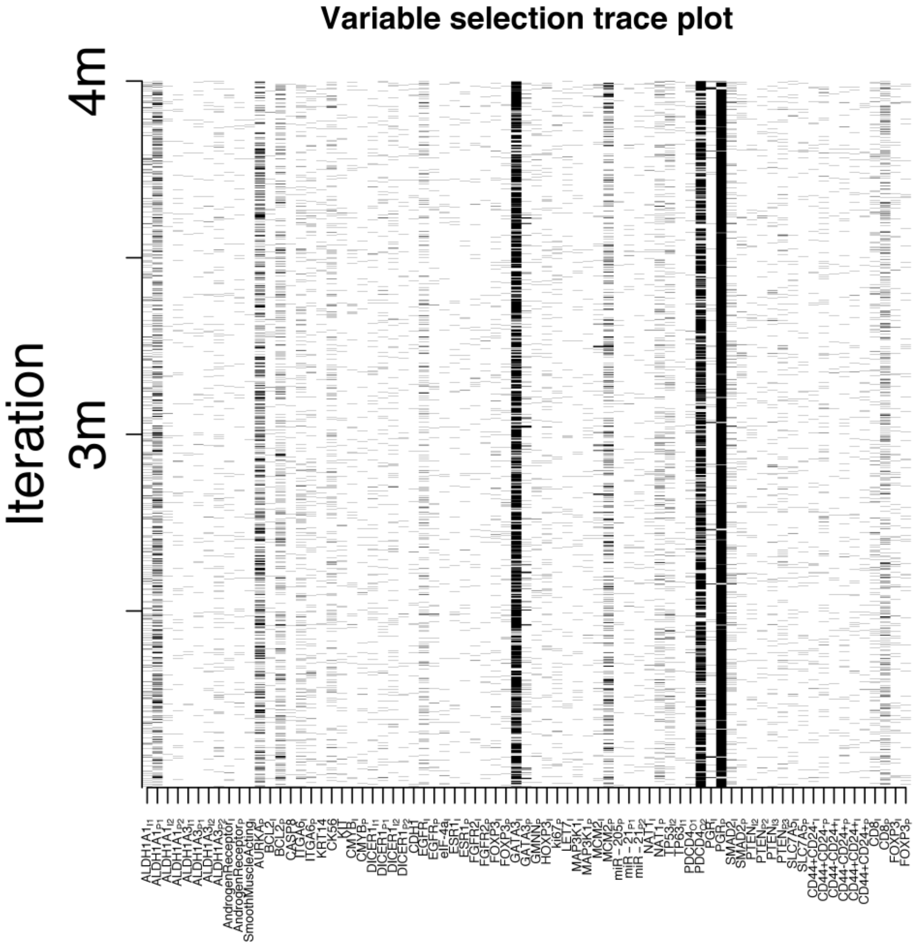

Figure S2: Trace plot of selection indicators provided by the SBWR analysis among predictor selections from the analysis of SEARCH. Predictors are ordered horizontally, and posterior samples from SBWR are ordered vertically from bottom to top. For each predictor, inclusion at the particular iteration is denoted in black, and exclusion is denoted in white. This plot helps to visualise the mixing patterns of the selection indicators and if variables seem to 'stick', i.e. do not come in and out in a fairly regular manner.

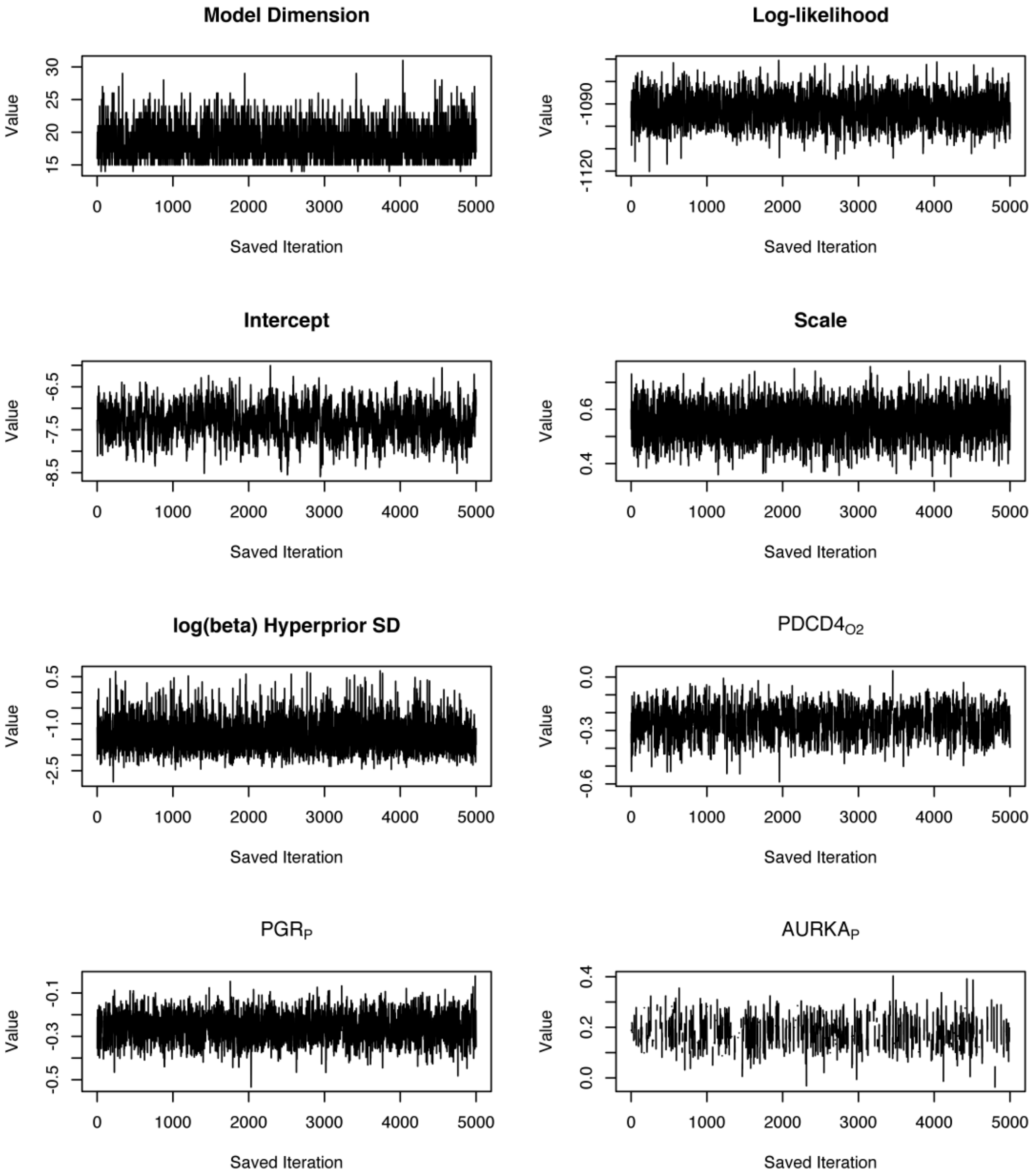

Figure S3: SBWR posterior chain plots from the analysis of SEARCH for a) the log-likelihood, b) the intercept α , c) the Weibull scale k, d) log(Beta) Hyperprior SD (σ_β) and the top three tumour markers e) - g).

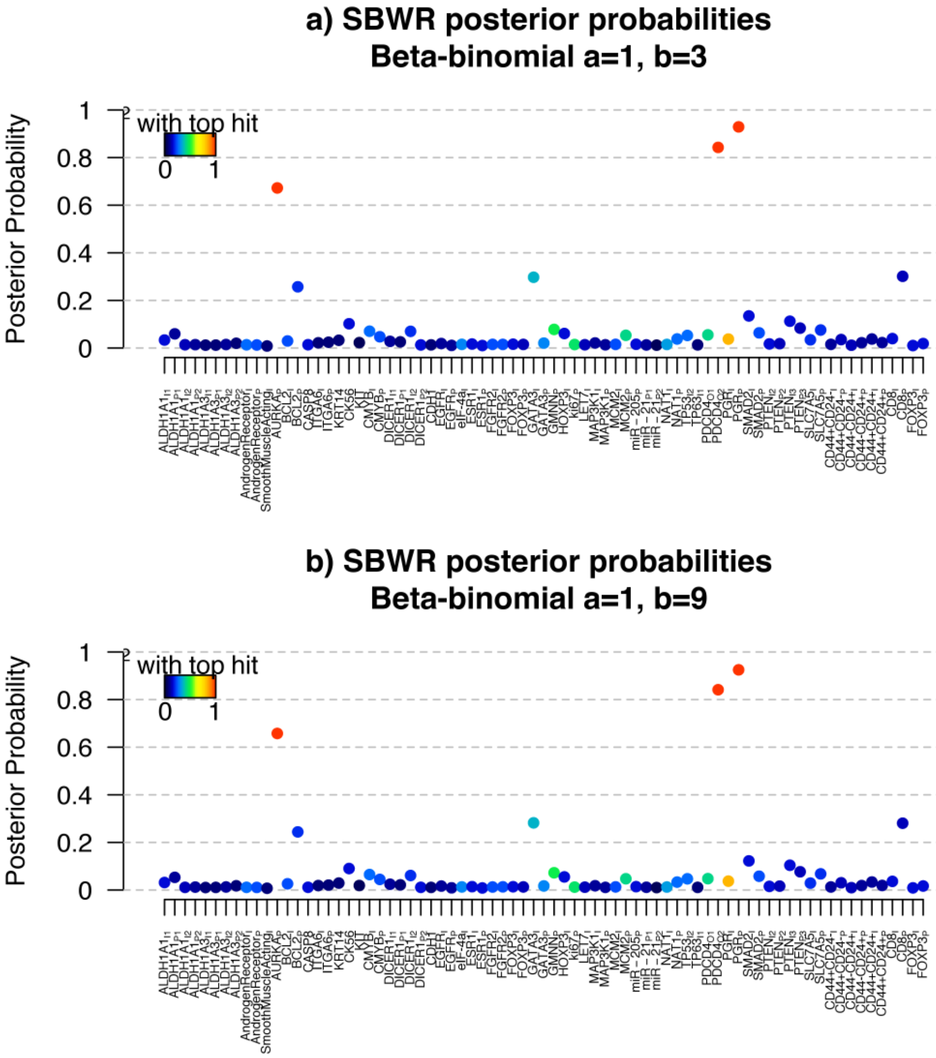

Figure S4: Sensitivity analyses of SBWR applied to the SEARCH data, for different choices of a_ω and a_ω in the Beta-binomial model space prior.

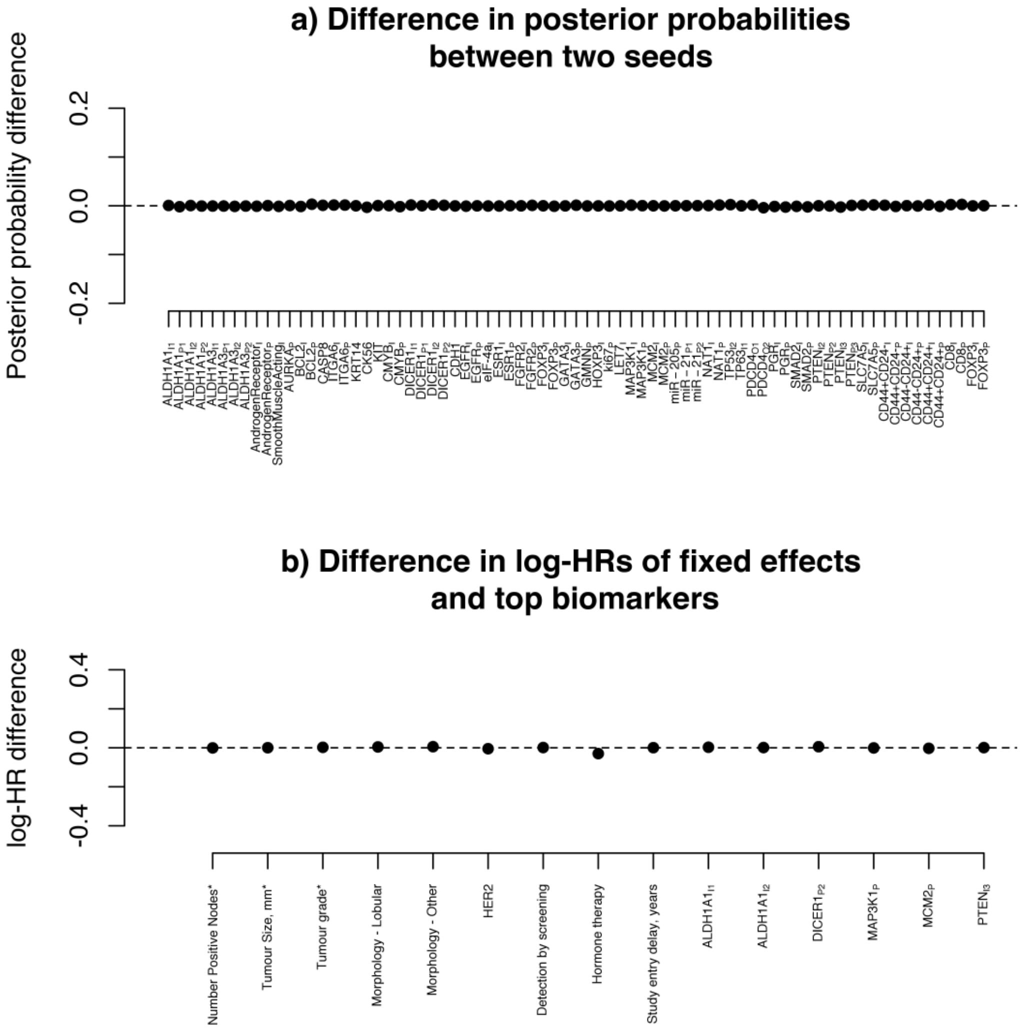

Figure S5: Comparison of SBWR posterior results from two different MCMC chains. Panel a) shows the difference in posterior probabilities inferred for each of the tumour markers between the chains. Panel b) Shows the difference between chains in median log-HR estimates for the fixed covariates, and for the tumour markers with posterior probabilities at least 25%. For the tumour markers, median log-HRs were calculated conditional on inclusion in the model.

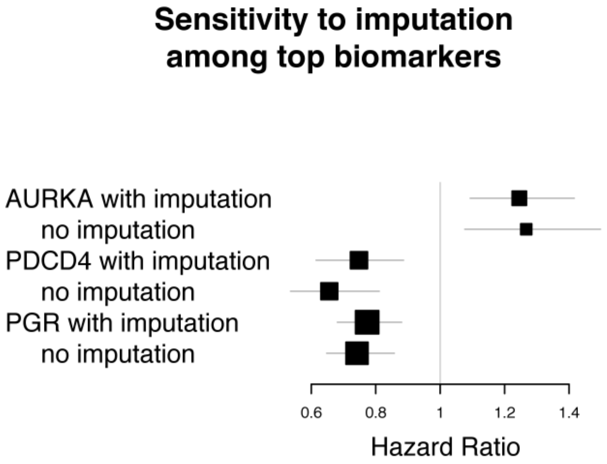

Figure S6: Exploration of sensitivity of SEARCH analysis results to imputation. For each of the top three tumour markers, the top box and interval represent the median Hazard Ratio and Credible Interval from the main SBWR analysis, in which 20 multiply imputed datasets were analysed. The lower box and interval represent results from SBWR univariate analyses in which imputed values were excluded for the particular tumour marker. These latter analyses were univariate since only a small number of patients had values observed for all three tumour markers simultaneously. These particular three tumour markers do not have strong correlations so univariate point estimates should be comparable to multivariate point estimates. The large overlap in the pairs of credible intervals for each tumour marker suggest the multiply imputed data is reasonable. It is to be expected that the point estimates from the analysis of imputed data shift slightly to the null. Evidence of association is generally weaker among imputed values due to the uncertainty associated with them. Therefore their incorporation into an analysis leads to some deterioration of the effect estimates.

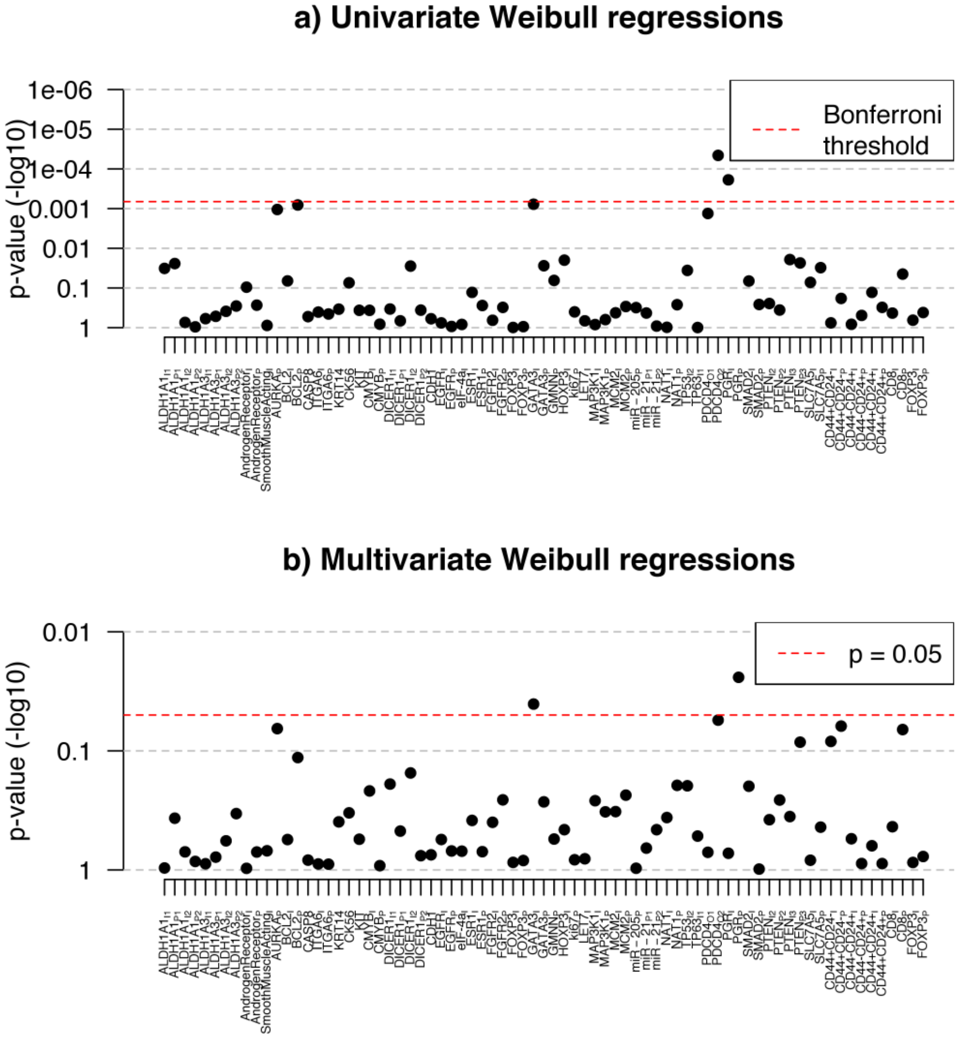

Figure S7: Results from standard Weibull regressions of the SEARCH data, in which p-values are combined across multiple imputation chains using Rubin's rules. The top panel displays univariate p-values of each tumour marker, and the bottom panel displays adjusted p-values in which the saturated model was fitted to the data from each multiple imputation chain.

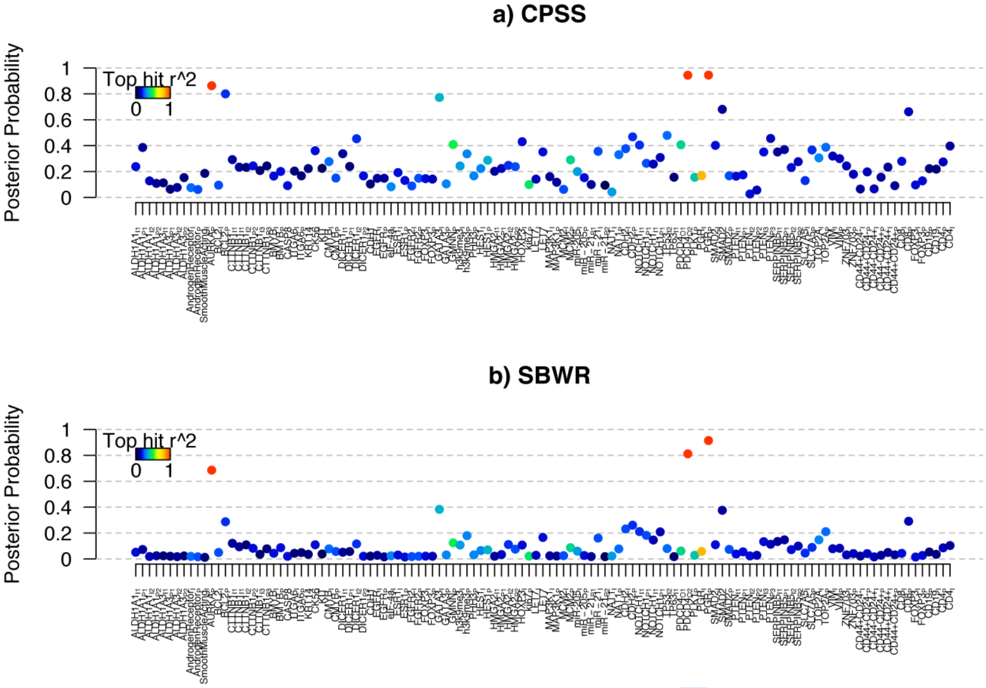

Figure S8: Association of tumour markers with breast cancer survival in the SEARCH dataset. All 119 tumour markers were included in these analyses. In both panels, each tumour marker is coloured according to its strongest pairwise correlation with one of the three top hits — PDCD4, PGR and AURKA. a) Selection probabilities from LASSO Cox regression with CPSS. b) SBWR posterior probabilities.