Hierarchical Bayesian Survival Analysis and Projective Covariate Selection in Cardiovascular Event Risk Prediction

Tomi Peltola
tomi.peltola@aalto.fi
Aalto University, Finland

Aki S. Havulinna
aki.havulinna@thl.fi
National Institute for Health and Welfare, Finland

Veikko Salomaa
veikko.salomaa@thl.fi
National Institute for Health and Welfare, Finland

Aki Vehtari
aki.vehtari@aalto.fi
Aalto University, Finland

Abstract

Identifying biomarkers with predictive value for disease risk stratification is an important task in epidemiology. This paper describes an application of Bayesian linear survival regression to model cardiovascular event risk in diabetic individuals with measurements available on 55 candidate biomarkers. We extend the survival model to include data from a larger set of non-diabetic individuals in an effort to increase the predictive performance for the diabetic subpopulation. We compare the Gaussian, Laplace and horseshoe shrinkage priors, and find that the last has the best predictive performance and shrinks strong predictors less than the others. We implement the projection predictive covariate selection approach of Dupuis and Robert (2003) to further search for small sets of predictive biomarkers that could provide cost-efficient prediction without significant loss in performance. In passing, we present a derivation of the projective covariate selection in Bayesian decision theoretic framework.

1 INTRODUCTION

Improving disease risk prediction is a major task in epidemiological research. Non-communicable diseases, many of which develop and progress slowly, are a major cause of morbidity worldwide. Accurate risk prediction could be used to screen individuals for targeted intervention. Advances in measurement technologies allow researchers cost-efficient quantification of large numbers of potentially relevant biomarkers, for example, in blood samples. However, often only a few of such candidate biomarkers could be expected to give practically relevant gain in risk stratification or could be realistically used in routine health care setting. The statistical challenge is then to identify an informative subset of the biomarkers and estimate its predictive performance.

Here, we describe an application of linear, hierarchical Bayesian survival regression to model cardiovascular event risk in diabetic individuals. The available data consists of 7932 Finnish individuals in the FINRISK 1997 cohort [1], of whom 401 had diabetes at the beginning of the study. The covariates consist of a set of 55 candidate biomarkers measured from blood samples and 12 established risk factors (e.g., baseline age, sex, body-mass index, lipoprotein cholesterol measures, blood pressure and smoking). The length of the follow-up period was 15 years. We focus on three key elements in the model construction: 1) using shrinkage priors to model the assumption of possibly limited relevance of many biomarkers, 2) utilizing the large set of non-diabetic individuals in the modelling, and 3) the selection of a subset of the biomarkers with predictive value. While the statistical approach is not limited to this particular application, we use the setting to make the description of the methods concrete.

Shrinkage or sparsity-promoting priors for regression coefficients are used to shrink the effects of (apparently) irrelevant covariates to zero, while retaining the effects of relevant covariates. Their use has increased with the availability of datasets with large numbers of features, for example, from high-throughput measurement technologies, which often capture a snapshot of a whole system (e.g., metabolome, genome) instead of targeted features. The interest has spawned considerable research effort into such priors and multiple alternatives have been proposed (see, e.g., refs [2–6]). In this work, we chose to compare three priors: the Laplace [3], the horseshoe [5] and, as a baseline, a Gaussian prior. The Laplace prior corresponds to the popular lasso penalty [7] in non-Bayesian regularized regression. The horseshoe prior has been shown to have desirable features in Bayesian analysis [5, 8]. We briefly review these priors in Section 2.2.

Of the 401 diabetic individuals in the study, 155 experienced a cardiovascular event within the follow-up period. This leaves a limited set of informative samples to perform the model fitting, covariate selection and predictive performance evaluation with. Although the risk of cardiovascular events is larger in diabetic individuals than the general population [9], we would expect that the risk factors are shared at least to some extent. Based on this assumption, we incorporate the non-diabetic individuals ($n = 7531$, 1031 events) into the analysis by constructing a hierarchical joint model, where the submodels for diabetic and non-diabetic individuals can be correlated (akin to transfer or multi-task learning [10]). The joint model does not place hard constraints on the similarity of the submodels, but allows the models to differ between non-diabetic and diabetic individuals and also between men and women. Details are given in Section 2.3.

While lasso regression in the non-Bayesian context can perform hard covariate selection by estimating exact zeroes for regression coefficients, the Bayesian shrinkage priors do not lead to sparse posterior distributions as there will remain uncertainty after observing a finite dataset. However, we are interested in finding a minimal subset of predictively relevant biomarkers as discussed above. To this end, we examine the use of projection predictive covariate selection[1], where the full model, encompassing all the candidate biomarkers and the uncertainties related to their effects, is taken as a *yardstick* for the smaller models. Specifically, the models with subsets of covariates are found by maximizing the similarity of their predictions to this reference as proposed by Dupuis and Robert [12]. Notably, this approach does not require specifying priors for the submodels and one can instead focus on building a good reference model. Dupuis and Robert [12] suggest choosing the size of the covariate subset based on an acceptable loss of explanatory power compared to the reference model. We examine using cross-validation based estimates of predictive performance as an alternative.

The structure of this article is as follows. In Section 2, we describe the survival model, shrinkage priors, and the hierarchical extension to include data of non-diabetic individuals. The projection predictive covariate selection is described in Section 3. The results from the application of the methods for cardiovascular-event-free survival modelling in diabetic individuals are presented in Section 4. Finally, Section 5 discusses the modelling approach.

[1] A comprehensive review of predictive Bayesian model selection approaches is given by Vehtari and Ojanen [11]. Our terminology follows theirs.

2 MODEL

We first consider modelling the cardiovascular-event-free survival in the subset of diabetic individuals only. The model is then extended to include the data of non-diabetic individuals, while allowing the covariate effects and the baseline hazard to differ in these groups and between men and women.

2.1 OBSERVATION MODEL

Let the observation t_i be the event time T_i or the censoring time C_i since the beginning of the study for ith individual and v_i be the corresponding event/censoring indicator (1 for observed events, 0 for censored). All censored cases are right censored (i.e., $T_i > C_i$ where only C_i is observed; censoring occurs in the data mostly because of event-free survival to the end of the follow-up). Further, let $\boldsymbol{x}_i$ be a column vector of the observed covariate values for the ith subject. We assume a parametric survival model, where the observations follow the Weibull model[2]

$$p(t_i|\boldsymbol{x}_i, v_i, \boldsymbol{\beta}, \alpha) = \alpha^{v_i} t_i^{v_i(\alpha-1)} \exp(v_i \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i - t_i^{\alpha} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i))$$

with the shape α and the scale defined through the linear combination $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i$ of the covariates [14]. The Weibull model is a proportional hazard model with the hazard function $h(T_i) = \alpha T_i^{\alpha-1} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i)$.

We include a constant term 1 in the covariates $\boldsymbol{x}_i$ and denote the corresponding regression coefficient β_0. The intercept and the shape are given the diffuse priors:

$$\begin{aligned} \beta_0 &\sim \mathrm{N}(0, 10^2), \\ \log \alpha &\sim \mathrm{N}(0, 10^2). \end{aligned}$$

The covariates are divided into a set of established risk (or protective) factors and a set of new candidate biomarkers, which are of more uncertain relevance. The coefficients of the established predictors, β_j for $j = 1, \ldots, m_{bg}$, are given the prior [15]:

$$\begin{aligned} \beta_j &\sim \mathrm{N}(0, \sigma_s^2 \sigma_j^2), \text{ for } j = 1, \ldots, m_{bg}, \\ \sigma_j^2 &\sim \text{Inv–}\chi^2(1), \text{ for } j = 1, \ldots, m_{bg}, \\ \sigma_s &\sim \text{Half–N}(0, 10^2). \end{aligned}$$

Priors for the coefficients of the candidate biomarkers are considered below.

[2] The notation for probability distributions follows the parametrizations given in ref. [13], except for the Weibull model, which is explicitly written out. *Half*-distributions refer to the restriction to the real positive axis.

2.2 PRIORS FOR BIOMARKER COEFFICIENTS

Based on our prior assumption that only some of the biomarkers are expected to be practically relevant for prediction, we consider the use of shrinkage priors for the biomarker coefficients. As discussed in the introduction, there has been a lot of recent research into these type of priors and there are multiple proposals. We restrict our consideration to three alternatives: the horseshoe prior [5], the Laplace prior [3], and, as a baseline approach, a Gaussian prior. Each of these can be expressed as normal scale mixtures

$$\beta_j \sim \mathrm{N}(0, \tau_s^2 \tau_j^2), \text{ for } j = m_{bg}+1, \ldots, m_{bg}+m_{bm},$$

where τ_s is a global scale parameter (shared across j) and τ_j are local parameters. Ideally, the prior shrinks the coefficients of irrelevant biomarkers to zero, but allows large coefficients for relevant biomarkers. In a sparse situation, with many irrelevant biomarkers and few relevant, this could be effected by making τ_s small, but allowing some τ_j to take on large values to escape the shrinkage [16].

The priors for τ_js, for $j = m_{bg}+1, \ldots, m_{bg}+m_{bm}$, for the three alternatives are

$$\begin{aligned} \tau_j &\sim \text{Half–Cauchy}(0,1) && \text{for horseshoe,} \\ \tau_j^2 &\sim \text{Exponential}(0.7) && \text{for Laplace,} \\ \tau_j &= 1 && \text{for the Gaussian.} \end{aligned}$$

A comparison of the Laplace and horseshoe prior is given in ref. [5]: it is noted that the Laplace prior may overshrink large coefficients in a sparse situation, while the horseshoe prior is more robust (see also ref. [16]). Furthermore, van der Pas et al. [8] derive theoretical results indicating that the posterior distribution under the horseshoe prior may be more informative[3] than under the Laplace prior in a sparse normal means problem. The Gaussian prior does not try to separate between relevant and irrelevant covariates as it depends only on the shared scale parameter τ_s.

The same prior is given for the global scale parameter in each case:

$$\tau_s \sim \text{Half–Cauchy}(0,1),$$

which has its (bounded) mode at zero, but is only weakly informative as it also places a substantial amount of prior mass far from zero (see refs [15–17] for discussion on priors for global variance parameters).

[3]That is, the posterior mean estimator attains a minimax risk, possibly up to a multiplicative constant, in a sparse setting and the posterior contracts at a similar rate (with conditions on τ_s).

2.3 HIERARCHICAL EXTENSION

Next, we consider extending the approach to jointly model the event-free survival of non-diabetic men (NM), non-diabetic women (NW), diabetic men (DM), and diabetic women (DW). Our aim is to increase the predictive performance of the model specifically in the subset of diabetic individuals, but gain power by including the larger set of observations for non-diabetic individuals in the model. To this end, we tie together the submodels of the four groups using the following assumptions:

1. The relevance of a biomarker will be similar for all the submodels.
2. The effect size of a biomarker (or other covariate) and its direction are similar between men and women, and between diabetic and non-diabetic individuals.
3. The baseline hazard functions have similar shapes for men and women, and diabetic and non-diabetic individuals.

Let $\boldsymbol{\beta}_j = [\beta_{j,NM}\ \beta_{j,NW}\ \beta_{j,DM}\ \beta_{j,DW}]^{\mathrm{T}}$ be the coefficients for the jth biomarker in the four submodels. We set

$$\boldsymbol{\beta}_j \sim \mathrm{N}(\mathbf{0}, r_j^2 \lambda \boldsymbol{\Lambda}^{-1}),$$

where $r_j^2 \lambda \boldsymbol{\Lambda}^{-1}$ is the prior covariance matrix. Here, $r_j = \tau_j \tau_s$ and follows one of the prior specifications given in the previous section. This encodes the first assumption above: a single r_j parameter defines the relevance of the jth biomarker in all the four submodels.

To encode the second assumption, we specify the structure of the prior precision matrix as

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1+c_N+s_M & -c_N & -s_M & 0 \\ -c_N & 1+c_N+s_W & 0 & -s_W \\ -s_M & 0 & 1+c_D+s_M & -c_D \\ 0 & -s_W & -c_D & 1+c_D+s_W \end{bmatrix}.$$

The corresponding graphical structure is illustrated in Figure 1. As will be made more explicit below, the c_N and c_D control the similarity of the submodels of non-diabetic men and women, and between the submodels of diabetic men and women, respectively. s_M and s_W control the similarity between the submodels of non-diabetic and diabetic men, and non-diabetic and diabetic women, respectively. We further simplify the model by taking $c_N = c_D = c$ and $s_M = s_W = s$ and constrain $c > 0$ and $s > 0$. The precision matrix has similarity to the one used by Liu et al. [18] to learn dependencies between covariates, but here $\boldsymbol{\Lambda}$ is restricted to encode a specific prior structure.

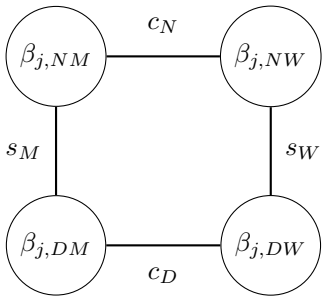

Figure 1: Prior structure for the regression coefficients of jth biomarker in the joint model.

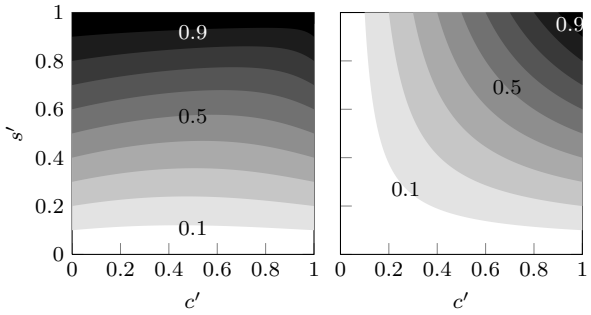

Figure 2: Contour plots of the correlation coefficient between $\beta_{j,NM}$ and $\beta_{j,DM}$ (left) and $\beta_{j,NM}$ and $\beta_{j,DW}$ (right) as a function of c' and s'.

We choose $\lambda = \frac{(2c+1)(2s+1)(2c+2s+1)}{(1+2c+2s+2cs)(c+s+1)}$ as this makes the diagonal elements of $\lambda\boldsymbol{\Lambda}^{-1}$ equal to 1, that is, $\lambda\boldsymbol{\Lambda}^{-1}$ becomes a correlation matrix. The relevance of the jth biomarker is then solely dependent on r_j.

For more insight, the prior for $\boldsymbol{\beta}_j$ may be written out as proportional to

$$\exp\left(-\frac{1}{2r_j^2\lambda}(S_2 + cS_c + sS_s)\right),$$

where $S_2 = \beta_{j,NM}^2 + \beta_{j,NW}^2 + \beta_{j,DM}^2 + \beta_{j,DW}^2$, $S_c = (\beta_{j,NM} - \beta_{j,NW})^2 + (\beta_{j,DM} - \beta_{j,DW})^2$ and $S_s = (\beta_{j,NM} - \beta_{j,DM})^2 + (\beta_{j,NW} - \beta_{j,DW})^2$. c controls the penalization in the difference between men and women, and s controls the penalization in the difference between non-diabetic and diabetic subjects. Taking negative logarithm of the prior shows that it corresponds to a specific Bayesian version of the multi-task graph regularization penalty proposed by Evgeniou et al. [19] and further studied by Sheldon [20]. The prior can also be represented in the sparse Bayesian multi-task learning framework of Archambeau et al. [21], where a zero-mean matrix-variate Gaussian density is placed on $\boldsymbol{B} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_m]$ with row covariance $\boldsymbol{\Omega}$ (over the m covariates) and column covariance $\boldsymbol{\Sigma}$ (over the *tasks*). Here, $\boldsymbol{\Omega}$ is a diagonal matrix with elements r_j^2 and $\boldsymbol{\Sigma} = \lambda\boldsymbol{\Lambda}^{-1}$.

We use the following transformations of c and s: $c = (1-c')^{-1} - 1$ and $s = (1-s')^{-1} - 1$, where $c' \in [0, 1)$ and $s' \in [0, 1)$. At $c' = 0$, $c = 0$ and the corresponding submodels are independent. As $c' \to 1$, $c \to \infty$ and the corresponding submodels are constrained to identical. s' behaves similarly.

We can also examine the implied prior distribution of the difference between two $\beta_{X,j}$ coefficients as a function of c' and s'. First, note that the distribution of $\beta_{X,j} - \beta_{Y,j}$ is $\mathrm{N}(0, 2r_j^2(1-\rho))$, where ρ is the correlation coefficient. Specifically, the variance of the distribution is linearly dependent on ρ and, for $\rho \geq 0$, has the maximum value of $2r_j^2$ when $\rho = 0$ and the minimum value of 0 when $\rho = 1$. In Figure 2, the implied prior correlation coefficients of some interesting pairs of $\beta_{X,j}$s are shown as functions of c' and s': s' controls almost linearly the correlation between $\beta_{j,NM}$ and $\beta_{j,DM}$, whereas the correlation between $\beta_{j,NM}$ and $\beta_{j,DW}$ is close to bilinear in c' and s'.

To complete the prior specification c' and s' are given prior distributions. We use different parameters for biomarkers (c' and s'), other covariates (c'_{bg} and s'_{bg}) and the log-scale Weibull shape parameter $\log\alpha$ (c'_α and s'_α; this encodes the third assumption):

$$\begin{aligned}
c' &\sim \mathrm{Beta}(a_c, b_c),\\
s' &\sim \mathrm{Beta}(a_s, b_s),\\
c'_{bg} &\sim \mathrm{Beta}(a_c, b_c),\\
s'_{bg} &\sim \mathrm{Beta}(a_s, b_s),\\
c'_\alpha &\sim \mathrm{Beta}(a_c, b_c),\\
s'_\alpha &\sim \mathrm{Beta}(a_s, b_s).
\end{aligned}$$

Finally, a_c, b_c, a_s and b_s are given $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{4})$ priors.

We note that the eigendecomposition of $\boldsymbol{\Lambda} = \boldsymbol{V}\boldsymbol{D}\boldsymbol{V}^{\mathrm{T}}$ is of simple form, with $\boldsymbol{D}$ being a diagonal matrix with elements 1, $1+2c$, $1+2s$, $1+2c+2s$ and

$$\boldsymbol{V} = \frac{1}{2}\begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

This can be useful in reparametrizing the model for Markov chain Monte Carlo sampling algorithms. It also shows that the precision matrix is positive definite.

3 METHODS FOR BIOMARKER SELECTION AND PREDICTIVE PERFORMANCE EVALUATION

The approaches used for biomarker selection and evaluation of predictive performance are described below. The model constructed in previous section is used as the reference model in the biomarker selection.

3.1 PROJECTION PREDICTIVE COVARIATE SELECTION

Assuming the availability of a reference model, which is a good representation of the predictive power of the candidate biomarkers and the related uncertainty, we seek a subset of the biomarkers, which can be used for prediction without a large loss in performance relative to the reference model. Our prior assumption of sparsity in the biomarker effects implies that this goal could be achievable. We describe the approach in two steps: 1) defining a submodel for making predictions with a specific subset of the candidate biomarkers, and 2) finding submodels with good predictive performance.

3.1.1 Projective Submodels

We use the projective approach of Dupuis and Robert [12], Goutis and Robert [22] to find the parameters of the submodel, but present an alternative derivation in the Bayesian decision theoretic framework reviewed in ref. [11]. The *projection* is posed as a solution to an optimization problem with regard to a restriction of the reference model. Let the covariates $\boldsymbol{x}$ be divided into two parts $\boldsymbol{x} = [\boldsymbol{x}_\perp, \boldsymbol{x}_\top]$ and define a submodel $M_\perp$ to be restricted to using the covariates in $\boldsymbol{x}_\perp$[4] with parameters $\boldsymbol{\theta}_\perp = (\boldsymbol{\beta}_\perp, \alpha_\perp)$ in the Weibull model. We find the submodel by maximizing the Gibbs reference utility

$$\bar{u}(M_\perp) = \int \left[\int u(M_\perp, \boldsymbol{x}_\perp, \boldsymbol{\theta}, T) p(T|\boldsymbol{\theta}, \boldsymbol{x}) dT \right] p(\boldsymbol{\theta}|D) p(\boldsymbol{x}) d(\boldsymbol{\theta}, \boldsymbol{x})$$

with respect to the unknown probability densities $f(\boldsymbol{\theta}_\perp|\boldsymbol{\theta})$ appearing in the $u(M_\perp, \boldsymbol{x}_\perp, \boldsymbol{\theta}, T) = \int f(\boldsymbol{\theta}_\perp|\boldsymbol{\theta}) \log p(T|\boldsymbol{\theta}_\perp, \boldsymbol{x}_\perp) d\boldsymbol{\theta}_\perp$. Here, $p(\boldsymbol{\theta}|D)$ is the posterior distribution of the reference model given the observed data D and $p(\boldsymbol{x})$ is the distribution of the covariates. Writing out u and changing the integration order,

$$\bar{u}(M_\perp) = \int \left[\int p(T|\boldsymbol{\theta}, \boldsymbol{x}) \log p(T|\boldsymbol{\theta}_\perp, \boldsymbol{x}_\perp) dT \right] \times f(\boldsymbol{\theta}_\perp|\boldsymbol{\theta}) p(\boldsymbol{\theta}|D) p(\boldsymbol{x}) d(\boldsymbol{\theta}_\perp, \boldsymbol{\theta}, \boldsymbol{x}).$$

Finally, to arrive at the same solution with Dupuis and Robert [12], $f(\boldsymbol{\theta}_\perp|\boldsymbol{\theta})$ can be restricted to the Dirac delta function $\delta(\boldsymbol{\theta}_\perp - \hat{\boldsymbol{\theta}}_\perp)$ with an offset $\hat{\boldsymbol{\theta}}_\perp$ that depends on $\boldsymbol{\theta}$. That is, the solution to the maximization of $\bar{u}$ is defined pointwise for each $\boldsymbol{\theta}$ as the corresponding optimal value of $\hat{\boldsymbol{\theta}}_\perp$. The pointwise solution arises from the dependence of f on $\boldsymbol{\theta}$.

As $p(\boldsymbol{\theta}|D)$ is not available analytically and $p(\boldsymbol{x})$ at all, the former is approximated with Markov chain Monte Carlo methods and the latter by using $\boldsymbol{x}_i$ samples available in the data D [12]. The obtained estimate is

$$\bar{u}(M_\perp) \approx \frac{1}{nJ} \sum_{i,j} \left[\int p(T|\boldsymbol{\theta}^{(j)}, \boldsymbol{x}_i) \log p(T|\hat{\boldsymbol{\theta}}_\perp^{(j)}, \boldsymbol{x}_{i,\perp}) dT \right],$$

where the double sum runs over the n data points and the J posterior samples. The optimization problems to find the optimal $\hat{\boldsymbol{\theta}}_\perp^{(j)}$s are independent over j. We solve them using the Newton's method.

We define the projection predictive distribution for the submodel $M_\perp$ as

$$p(T|\boldsymbol{x}_\perp, M_{ref}) = \int p(T|\boldsymbol{x}_\perp, \boldsymbol{\theta}_\perp) f(\boldsymbol{\theta}_\perp|M_{ref}) d\boldsymbol{\theta}_\perp,$$

where we explicitly emphasize the dependence on the reference model M_{ref} and which is approximated using the projected samples $\hat{\boldsymbol{\theta}}_\perp^{(j)}$s. This kind of projected predictive distribution was also considered by Nott and Leng [23].

Note that scaling the estimated $\bar{u}$ as $d(M_\perp) = \bar{u}(M_{ref}) - \bar{u}(M_\perp)$ (and minimizing instead of maximizing) does not change the optimal solution and gives otherwise the same formula as $\bar{u}$, except the term in square brackets is replaced with the Kullback–Leibler divergence between $p(T|\boldsymbol{x}, \boldsymbol{\theta})$ and $p(T|\boldsymbol{x}_\perp, \boldsymbol{\theta}_\perp)$. This gives the approach further information theoretic justification and is the basis of the formulation in Dupuis and Robert [12]. They also suggest defining the relative explanatory power of the submodel as

$$\text{relative explanatory power}(M_\perp) = 1 - \frac{d(M_\perp)}{d(M_0)},$$

where M_0 refers to the model without any of the candidate biomarkers and which transforms the $d(M_\perp)$ values to between 0 (for $M_\perp = M_0$) and 1 (for $M_\perp = M_{ref}$).

[4] We assume that the established risk factors are always included in this set.

3.1.2 Submodel Search

$\bar{u}$ (or equivalently d) is used to compare the submodels in the search for good subsets of biomarkers. However, exhaustive search of the model space[5] is not feasible, unless the number of candidate biomarkers is small. We choose to use the suboptimal forward selection strategy for its simplicity and its scalability to large covariate sets:

1. Begin with the submodel M_0 (no biomarkers) and set j to 0.
2. Repeat until all biomarkers have been added:
   (a) Find the projections for all submodels that are obtainable by adding one new biomarker to M_j. Select the one with largest $\bar{u}$ and set it as M_{j+1}. Set j to $j+1$.

This defines a deterministic[6] path of models from M_0 to $M_{m_{bm}}$ and gives a ranking of the biomarkers according to their projection predictive value. Dupuis and Robert [12] suggest finally choosing the smallest submodel with an acceptable loss in the explanatory power relative to the reference model (and use a slightly more elaborate search). Alternatively, one could monitor some other statistic (e.g., predictive performance) along the search path to locate good submodels. Computing the full forward selection path may not be necessary, if a suitable stopping criterion is used in the step 2 above.

3.2 PREDICTIVE PERFORMANCE EVALUATION

Given a model M with posterior predictive distribution $p(T_*|\boldsymbol{x}_*, D)$, where D is the observed data, we evaluate its predictive performance using the logarithm of the predictive density (LPD) at an actual observation $(t_*, v_*, \boldsymbol{x}_*)$. This scoring rule is proper and measures the calibration and sharpness of the predictive distribution simultaneously [24]. As the predictive densities are not available analytically for the models considered here, we estimate the LPD score from the Markov chain Monte Carlo samples of the posterior distribution:

$$\mathrm{LPD}_*(M) \approx \log \frac{1}{J} \sum_j^J p(t_*|\boldsymbol{x}_*, v_*, \boldsymbol{\beta}^{(j)}, \alpha^{(j)}),$$

where $(\boldsymbol{\beta}^{(j)}, \alpha^{(j)})$ are J posterior samples of the model given the data D.

[5]The number of subsets for m_{bm} covariates is $2^{m_{bm}}$.

[6]Given the stochastic samples from the posterior distribution of the reference model.

Stratified ten-fold cross-validation [25] is used to obtain estimates of the generalization performance: The full dataset is divided randomly into ten disjoint subsets (folds), while balancing the sets to have approximately similar age distributions and proportions of diabetic and non-diabetic individuals, men and women, and cases of cardiovascular events. Predictions for each fold are obtained using a posterior distribution based on training data, where the particular fold has been left out. Given predictions obtained this way, the predictive performance is summarized by the mean LPD over the full set of n data points (MLPD).

To reduce variance and gauge uncertainty in model comparisons, we compute Bayesian bootstrap [26] samples of the MLDP difference (ΔMLPD) between model M_a and model M_b by

$$\Delta\mathrm{MLPD}^{(j)}(M_a, M_b) = \sum_{i=1}^{n} w_i^{(j)} [\mathrm{LPD}_i(M_a) - \mathrm{LPD}_i(M_b)],$$

where $w_i^{(j)}$, $i = 1, \ldots, n$, are the bootstrap weights ($\sum_i w_i^{(j)} = 1$) for the jth bootstrap sample generated using the Dirichlet distribution with parameters set to 1 [11]. The comparison is summarized by the q-value[7]:

$$q(M_a, M_b) = \frac{1}{J} \sum_{j=1}^{J} I(\Delta\mathrm{MLPD}^{(j)} \geq 0),$$

where $I(\cdot) = 1$ if the given condition holds and 0 otherwise, and which is interpreted as the Bayesian posterior probability (under the Dirichlet model) of M_a performing better than M_b [11].

4 RESULTS

Missing values in the covariate data were multiply imputed using chained linear regressions with in-house scripts based on ref. [27]. The candidate biomarkers were log-transformed and scaled to have zero mean and unit variance. The No-U-Turn variant of the Hamiltonian Monte Carlo algorithm [28], as implemented in Stan software [29], was used to sample from the posterior distributions of the full models. The sampling was done independently for 5 imputed datasets (4 chains of 1000 samples after burn-in for each). The samples were then concatenated. The sampling process was further performed independently for each of the 10 cross-validation training sets. All shown estimates of predictive performance were computed using cross-validation (Section 3.2).

[7]We use q instead of p to avoid confusion with the frequentist p-value.

Table 1: Model comparisons on cross-validation predictions. MLPDs and q-values (Section 3.2) are shown for predictions only on diabetic women, only on diabetic men or both. q-values are calculated against the *joint horseshoe* model; color scale 0.0 1.0.

	women		men		women & men	
model	MLPD	q-value	MLPD	q-value	MLPD	q-value
joint horseshoe	-0.581	NA	-0.716	NA	-0.652	NA
joint Laplace	-0.582	0.27	-0.720	0.10	-0.656	0.08
joint Gaussian	-0.585	0.22	-0.727	0.05	-0.660	0.04
joint no-biomarkers	-0.594	0.18	-0.758	0.03	-0.681	0.01
diab women&men horseshoe	-0.606	0.03	-0.719	0.44	-0.666	0.13
diab women/men horseshoe	-0.610	0.03	-0.721	0.45	-0.669	0.15
diab women/men no-biomarkers	-0.613	0.05	-0.765	0.04	-0.694	0.01

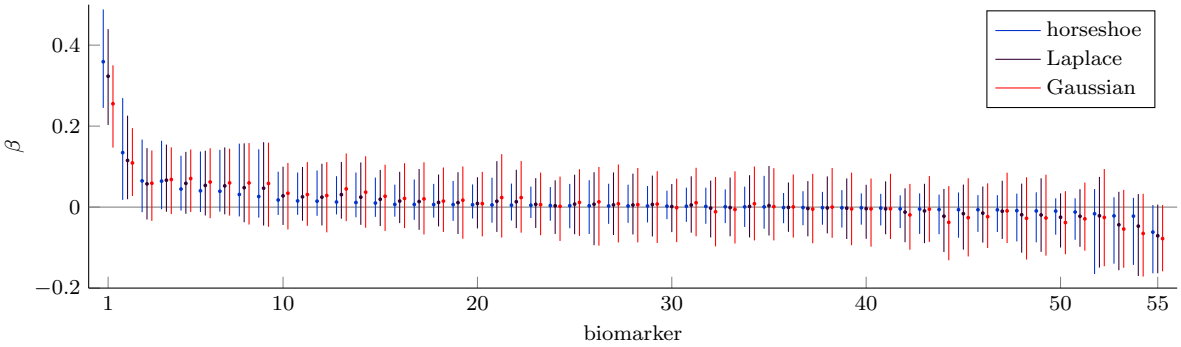

Figure 3: Biomarker regression coefficients β for the submodel of diabetic men in the joint models with the horseshoe, Laplace and Gaussian priors (full dataset). Dot is the mean and vertical line shows the 95% credible interval. Biomarkers are ordered according to the mean coefficients of the horseshoe model.

4.1 MODEL COMPARISONS

Table 1 presents results on comparing the mean log predictive densities (MLPD) of the following combinations of models: *joint* for the joint model of non-diabetic and diabetic individuals (Section 2.3), *diab women&men* for a joint model of diabetic men and women (two-group version of Section 2.3), *diab women/men* for separate models of diabetic men and women (without the extension of Section 2.3), and using the *horseshoe*, *Laplace* or *Gaussian* priors on the biomarker effects, or using only the established risk factors (*no-biomarkers*). The MLPDs and q-values were computed separately for the predictions for women and men, and for pooled predictions, and, importantly, in each case only for the predictions on the diabetic subpopulation.

The results show that there is an increase in the predictive performance when supplanting the established risk factors with the candidate biomarkers. The increase holds both when using the joint models or using only the data of diabetic individuals and seems to be greater in men. This indicates that the candidate biomarkers contain relevant information for predicting cardiovascular event risk.

Including the data of the non-diabetic individuals in the model seems to increase the predictive performance for the diabetic subpopulation, especially for women. The covariate effects in the joint models are very similar across the diabetic and non-diabetic submodels: posterior mean of s' is 0.96 for the horseshoe model. This implies that the risk factors behave similarly in both groups, but it is also possible that the dataset has limited information to distinguish between them and that larger datasets could uncover more differences.

Finally, it seems that the horseshoe prior performs better than the Laplace, and that the Gaussian is the worst of the three for this data. Figure 3 shows a comparison of the biomarker regression coefficients under these priors. The Laplace and the Gaussian priors

shrink the largest coefficient more than the horseshoe as would be expected in a sparse setting [5, 16]. Furthermore, the horseshoe seems to shrink coefficients near zero more strongly than the Laplace making the credible intervals around zero narrower.

4.2 BIOMARKER SELECTION AND SUBMODEL PREDICTIVE PERFORMANCE

We applied the projection predictive covariate selection (Section 3.1) with the joint horseshoe model as the reference. The forward selection was run using only the part of the model concerning diabetic individuals. We run the forward selection jointly for women and men to get an overall biomarker ranking for the diabetic subpopulation. The forward selection was run also for each cross-validation training set separately (using the reference model fitted on the corresponding training data).

Figure 4 shows the relative explanatory power curves along the forward selection path. In the full dataset, the best candidate biomarker attains 61% explanatory power relative to the reference model, five best reach over 80% and ten biomarkers are needed to reach over 90%. The growth in the explanatory power slows with more biomarkers, indicating diminishing gains from adding more candidate biomarkers (22 are needed to reach 95% and the remaining 33 account for the last 5%).

However, choosing an acceptable loss in the explanatory power to select an appropriate minimal subset of the biomarkers for use in prediction tasks seems difficult. In Figure 5, we show MLPDs (normalized to the reference model) obtained using the projection predictive covariate selection approach within the cross-validation. Top panel shows the ΔMLPD along the forward selection path and the bottom panel by the obtained relative explanatory power (e.g., at 0.6, the predictions in each cross-validation fold was made with the smallest submodel reaching 60% power in that fold). These show a mode at 2 biomarkers and at around 0.65 relative explanatory power (which corresponds to choosing two, three or four biomarkers depending on the fold). A second peak can be seen at 10 biomarkers or correspondingly at 0.91 power (10–16 biomarkers).

Unfortunately, the variance in the cross-validation estimates is quite large for making a definite choice based on them. Figure 6 shows the full set of pairwise comparisons between the submodels along the forward selection path (by number of biomarkers; same as in Figure 5 top panel). This indicates that two biomarkers is overall the best choice, but the difference to the 10

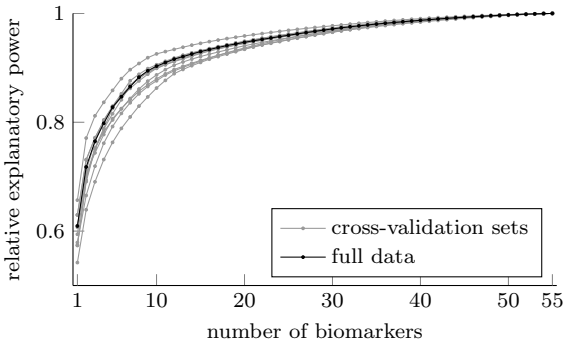

Figure 4: Relative explanatory powers along the forward selection path.

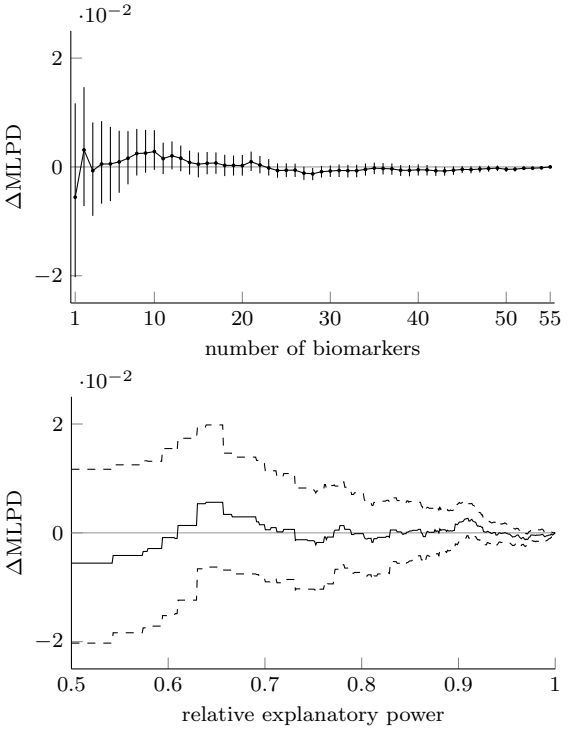

Figure 5: ΔMLPD values (in reference to the full model) by number of biomarkers (top) or by explanatory power threshold (bottom). Top: vertical lines are 95% Bayesian bootstrap credible intervals for the ΔMLPD. Bottom: dashed curves show the (pointwise) credible interval.

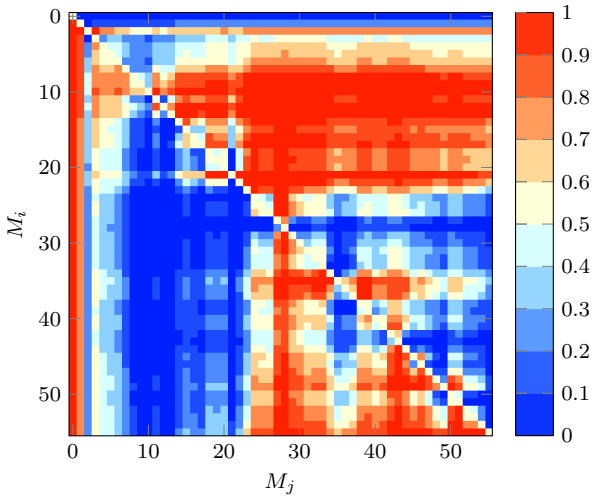

Figure 6: q-value matrix for M_i is better than M_j with regard to MLPD (Section 3.2; M.s refer to the submodels along the forward selection path).

biomarker selection is not large (q-value = 0.52). However, on comparing these to the full model or generally models with 11 or more biomarkers, the 10 biomarker selection is more confidently better (q-values mostly > 0.9) than the 2 biomarker selection (q-values mostly within 0.7–0.8).

Nevertheless, the analysis seems to support two clearly predictively relevant biomarkers for the cardiovascular risk prediction, with further 8 possibly interesting candidate biomarkers, but with some uncertainty about their relevance. Figure 3 also supports this conclusion with two of the biomarkers having clearly non-zero effects.

5 DISCUSSION

This paper presented a Bayesian analysis of cardiovascular-event-free survival in diabetic individuals, with the aim of identifying biomarkers with predictive value. We presented a comparison of the horseshoe, Laplace and Gaussian priors on the candidate biomarker effects and demonstrated empirically an expected [5, 16] difference in their behaviour. We further extended the model hierarchically to include data of non-diabetic individuals and examined the use of projection predictive covariate selection to find biomarker subsets with good predictive performance.

We could also hope that the predictive biomarkers capture some part of the state of the underlying disease process and as such could be used to speculate about causal disease pathways and to prioritize biomarkers for further study. However, the analysis approach does not warrant any formal causal inferences. Moreover, the inclusion of the data of non-diabetic individuals may bias the inferences on the diabetic subpopulation towards the general population, when the dataset has limited information to distinguish them. Nevertheless, the presented predictive comparisons, being independent of the model assumptions, justify studying the joint model.

The submodels in projection predictive covariate selection depend on the observed data only through the reference model. Thus, finding the submodel parameters and the covariate selection itself do not cause further fitting to the data, but rely on the information provided by the reference model [11]. The projected submodels may also be able to retain some predictive features of the reference model that would not be available, if the submodels were independently fitted to the data [11]: importantly, from Bayesian point of view, the submodel may be able to account for uncertainty due to the omission of some covariates.

However, selecting a single submodel for future prediction tasks may be difficult. We examined using the projection approach within cross-validation to obtain estimates of the submodel predictive performances. A disadvantage of this procedure is that the performance estimates are for the selection process and not for some particular combination of selected biomarkers. Furthermore, if selection is based on these estimates, the performance estimate for the chosen submodel will not anymore be unbiased for out-of-sample prediction unless nested cross-validation is used [11].

Acknowledgements

We acknowledge the computational resources provided by Aalto Science-IT project.

References

[1] Erkki Vartiainen, Tiina Laatikainen, Markku Peltonen, Anne Juolevi, Satu Männistö, Jouko Sundvall, Pekka Jousilahti, Veikko Salomaa, Liisa Valsta, and Pekka Puska. Thirty-five-year trends in cardiovascular risk factors in Finland. *International Journal of Epidemiology*, 39(2):504–518, 2010.

[2] T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404): 1023–1032, 1988.

[3] Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[4] Jim E. Griffin and Philip J. Brown. Inference

with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

[5] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[6] Zhihua Zhang, Shusen Wang, Dehua Liu, and Michael I. Jordan. EP-GIG priors and applications in Bayesian sparse learning. *The Journal of Machine Learning Research*, 13(1):2031–2061, 2012.

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

[8] S. L. van der Pas, B. J. K. Kleijn, and A. W. van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *arXiv preprint arXiv:1404.0202*, 2014.

[9] Emerging Risk Factors Collaboration. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet*, 375(9733):2215–2222, 2010.

[10] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[11] Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[12] Jérome A. Dupuis and Christian P. Robert. Variable selection in qualitative models via an entropic explanatory power. *Journal of Statistical Planning and Inference*, 111(1):77–94, 2003.

[13] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. CRC press, third edition, 2014.

[14] Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer, 2001.

[15] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

[16] Nicholas G. Polson and James G. Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 9*, pages 501–538. Oxford University Press, 2011.

[17] Nicholas G. Polson and James G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012.

[18] Fei Liu, Sounak Chakraborty, Fan Li, Yan Liu, and Aurelie C. Lozano. Bayesian regularization via graph Laplacian. *Bayesian Analysis*, 9(2): 449–474, 2014.

[19] Theodoros Evgeniou, Charles A. Micchelli, and Massimiliano Pontil. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

[20] Daniel Sheldon. Graphical multi-task learning. In *NIPS 2008 Workshop: "Structured Input – Structured Output"*, 2008.

[21] Cédric Archambeau, Shengbo Guo, and Onno Zoeter. Sparse Bayesian multi-task learning. In *Advances in Neural Information Processing Systems 24*, pages 1755–1763, 2011.

[22] Constantinos Goutis and Christian P. Robert. Model choice in generalised linear models: A Bayesian approach via Kullback–Leibler projections. *Biometrika*, 85(1):29–37, 1998.

[23] David J. Nott and Chenlei Leng. Bayesian projection approaches to variable selection in generalized linear models. *Computational Statistics & Data Analysis*, 54(12):3227–3241, 2010.

[24] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[25] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1137–1145, 1995.

[26] Donald B. Rubin. The Bayesian bootstrap. *The Annals of Statistics*, 9(1):130–134, 1981.

[27] Stef van Buuren and Karin Groothuis-Oudshoorn. MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 2011.

[28] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014.

[29] Stan Development Team. Stan: A C++ library for probability and sampling, version 2.2, 2014. URL `http://mc-stan.org/`.