Integration of Survival Data from Multiple Studies

Steffen Ventz[$,*], Rahul Mazumder[#], Lorenzo Trippa[$]

[$]Harvard T.H. Chan School of Public Health and Dana-Farber Cancer Institute

[#]Massachusetts Institute of Technology

July 20, 2020

Abstract

We introduce a statistical procedure that integrates survival data from multiple biomedical studies, to improve the accuracy of predictions of survival or other events, based on individual clinical and genomic profiles, compared to models developed leveraging only a single study or meta-analytic methods. The method accounts for potential differences in the relation between predictors and outcomes across studies, due to distinct patient populations, treatments and technologies to measure outcomes and biomarkers. These differences are modeled explicitly with study-specific parameters. We use hierarchical regularization to shrink the study-specific parameters towards each other and to borrow information across studies. Shrinkage of the study-specific parameters is controlled by a similarity matrix, which summarizes differences and similarities of the relations between covariates and outcomes across studies. We illustrate the method in a simulation study and using a collection of gene-expression datasets in ovarian cancer. We show that the proposed model increases the accuracy of survival prediction compared to alternative meta-analytic methods.

[*]steffen.ventz.81@gmail.com

Keywords: Meta-Analysis, Survival Analysis, Risk Prediction, Hierarchical Regularization.

1 Introduction

Various biomedical technologies enable the use of omics information for prognostic purposes, to quantify the risk of diseases, or to predict response to treatments. Risk stratification in oncology often utilizes a set of biomarkers to predict cancer progression or death within a time period. The number of covariates can exceed the sample size. This makes the identification of relevant genomic features for risk prediction and the development of accurate models challenging. Penalized regression and methods that utilize multiple datasets have been discussed in this context. Penalization methods enable parameter estimation and prediction when the number of predictors is large [46]. Meta-analyses [16] and integrated analyses [12] combine information from multiple studies for parameter estimation and prediction [24, 48]. These statistical procedures improve the estimation of parameters of interest with respect to single-study estimates if the covariate-outcome relations are similar across studies [40, 3, 51]. For instance Riester et al. [40], Bernau et al. [3] and Waldron et al. [51] showed that meta-analytic procedures tend to outperform the prediction accuracy of models developed using only a single study. But [40, 51, 47] also described heterogeneity of covariate effects across cancer studies, due to differences in assays, treatments and patient populations.

We introduce a model for the integrated analysis of a collection of datasets, with the aim of improving the accuracy of predictions, compared to single-study models and meta-analytic procedures. We use study-specific parameters in covariate-outcome regression models. These parameters are estimated borrowing information across studies with hierarchical regularization, which shrinks the study-specific regression coefficients towards each other. We use a squared $K \times K$ similarity matrix representative of differences and similarities of the covariates' effects across K studies. The matrix is used to estimate the study-specific models. The

regression parameters of each study are shrunken more towards the parameters of similar studies and less towards the remaining studies.

In previous work on integrative analyses Liu et al. [28] discussed Bayesian methods and variable selection for accelerated failure time models. Hierarchical normal models for multi-study gene expression analyses have been developed in [12, 11, 13], and Ma et al. [30, 31] studied penalized regression methods for integrative analyses, focusing on binary and accelerated failure-time outcome models. For case-control analyses with multiple datasets, Liu et al. [29] proposed an adaptive group-LASSO (agLASSO) procedure, and Cheng et al. [10] extended the approach using different regularization techniques. In the following sections we introduce a procedure which builds on the work that we mentioned. The procedure shrink study-specific parameters towards each other accounting for the degrees of similarity specific of each pair of studies.

2 The multi-study model

We consider K studies with time-to-event outcomes and predictors such as gene expression measurements. For each study $k = 1, \ldots, K$ the vector $\boldsymbol{Y}_k = \{Y_{k,i}\}_{i=1}^{n_k}$ indicates (possibly censored) survival times of n_k individuals and $\boldsymbol{C}_k = \{C_{k,i}\}_{i=1}^{n_k}$ denotes the vector of censoring variables, where $C_{k,i} = 1$ if $Y_{k,i}$ is an observed event time and it is zero if there is censoring at time $Y_{k,i}$. The vector $\boldsymbol{x}_{k,i} \in \mathbb{R}^p$ represents a set of p predictors and $\boldsymbol{X}_k = (\boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,n_k})^T$. Lastly, $\mathcal{D}_k = (\boldsymbol{Y}_k, \boldsymbol{C}_k, \boldsymbol{X}_k)$ indicates the data from study k and $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$ is a collection of studies.

We assume that failure times in each study k follow a proportional hazard model [14] with baseline survival function $S_k(\cdot)$ and study-specific coefficients $\boldsymbol{\beta}_k \in \mathbb{R}^p$. The approach that we will describe can be applied to alternative time to event models, for instance to accelerated failure time models [53] or accelerated hazards models [9], computations would only require minor modifications.

Inference is based on the Breslow's modification of the partial log-likelihood functions [14] for (possible tied) survival times,

$$l\left(\boldsymbol{\beta}_k, \mathcal{D}_k\right) = \sum_{\ell=1}^{m_k} \left\{ \tilde{\boldsymbol{x}}'_{k,\ell}\boldsymbol{\beta}_k - d_{k,\ell} \log\left(\sum_{i:Y_{k,i}\geq t_{k,\ell}} \exp\{\boldsymbol{x}'_{k,i}\boldsymbol{\beta}_k\} \right) \right\},$$

where $\left\{t_{k,\ell}\right\}_{\ell=1}^{m_k}$ are the m_k unique event times in study k, $d_{k,\ell}$ denotes the number of observed events at time $t_{k,\ell}$, $\tilde{\boldsymbol{x}}_{k,\ell} = \sum_i \boldsymbol{x}_{k,i} I(C_{k,i} = 1, Y_{k,i} = t_{k,\ell})$, for $\ell = 1, \ldots, m_k$, and $I(A)$ is the indicator function of the event A.

When the number of predictors exceeds the number of observed events a unique maximum partial-likelihood estimate does not exist and maximization of the regularized likelihood function $l\left(\boldsymbol{\beta}_k, \mathcal{D}_k\right) - R\left(\boldsymbol{\beta}_k\right)$ has been proposed [46, 34, 43] to obtain covariate effect estimates. Here $R\left(\boldsymbol{\beta}_k\right)$ is a non-negative function that equals zero when $\boldsymbol{\beta}_k = \mathbf{0}$. Popular approaches include the LASSO, ridge, elastic-net and the bridge penalties to name a few [25, 45, 46, 19, 52, 43]. We refer to [5, 6, 50, 3] for comparisons of regularization methods for predicting survival outcomes using genomic profiles.

As demonstrated in [3] studies, with nearly identical aims, often presents different joint distributions of predictors $\boldsymbol{x}_{k,i}$ and outcomes $Y_{k,i}$. Clusters of studies which correspond to different essays, patient populations, treatments, and study designs have been discussed [3, 47]. We introduce a model with study-specific parameters $\boldsymbol{\beta}_k$, and a latent parameter $\boldsymbol{\beta}_0$, which can be interpreted as the mean parameter across studies. Some studies will have similar vectors $\boldsymbol{\beta}_k$ due to similarities in the assays and patient populations, while other studies might be considerably different [3, 51]. We estimate the vectors $\boldsymbol{\beta}_k$ by borrowing information from studies $k' \neq k$ that are similar to study k. At the same time, studies k' that differ substantially from study k will have little influence on the estimation of $\boldsymbol{\beta}_k$. In different words, borrowing of information mirrors similarities and differences across studies. The latent parameter and study-level parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_K)$ are estimated using the

regularized likelihood

$$l_R(\boldsymbol{\beta}) = \sum_{k=1}^{K} l(\boldsymbol{\beta}_k, \mathcal{D}_k) - R_0(\boldsymbol{\beta}_0) - R_1(\boldsymbol{\beta}). \tag{1}$$

Here the parameters $\boldsymbol{\beta}_k$ can be interpreted as a noisy realization of $\boldsymbol{\beta}_0$, the average effect across studies. The non-negative function $R_0(\cdot)$ regularizes $\boldsymbol{\beta}_0$ and is zero when $\boldsymbol{\beta}_0 = \mathbf{0}$ (for example a lasso penalty). Similarly, the non-negative function $R_1(\cdot)$ is zero when $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K$ (see below for examples) and is used to borrow information across studies in the estimation of $\boldsymbol{\beta}$. In our applications in Sections 4 and 5 we will use $\widehat{\boldsymbol{\beta}}_0$ for risk predictions of patients in populations $k > K$ that are not represented in our collection of K studies, whereas for patients belonging to populations $k = 1, \ldots, K$, the estimate $\widehat{\boldsymbol{\beta}}_k$ can be directly used for risk predictions.

Penalized maximum likelihood estimates based on (1) have a Bayesian interpretation. See [45, 23, 37, 35] for a discussion on the relations between regularization methods and Bayesian analyses. Consider a Bayesian model for the unknown parameters $\boldsymbol{\beta}$, with prior probability $Pr(\boldsymbol{\beta}_0) \propto e^{-R_0(\boldsymbol{\beta}_0)}$ for the vector $\boldsymbol{\beta}_0$ and $Pr(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K | \boldsymbol{\beta}_0) \propto e^{-R_1(\boldsymbol{\beta})}$ for the study specific parameters conditionally on $\boldsymbol{\beta}_0$. The approximate posterior density of $\boldsymbol{\beta}$ with respect to the partial likelihood (see [44] for a formal justification) is proportional to

$$Pr_{PL}(\boldsymbol{\beta}|\mathcal{D}) \propto Pr(\boldsymbol{\beta}_0) Pr(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K | \boldsymbol{\beta}_0) \prod_{k=1}^{K} e^{l(\boldsymbol{\beta}_k, \mathcal{D}_k)}. \tag{2}$$

Therefore the mode of (2) coincides with the parameter $\boldsymbol{\beta}$ that maximizes (1). If we set $R_1(\boldsymbol{\beta}) = \sum_k \tilde{R}_1(\boldsymbol{\beta}_k, \boldsymbol{\beta}_0)$, with $\tilde{R}_1(\boldsymbol{\beta}_k, \boldsymbol{\beta}_0) \geq 0$ then the Bayesian model (2) incorporates the assumption that studies are exchangeable with, conditionally on $\boldsymbol{\beta}_0$, independent and identically distributed covariate effects $\boldsymbol{\beta}_k$. For example $\tilde{R}_1(\boldsymbol{\beta}_k, \boldsymbol{\beta}_0) = ||\boldsymbol{\beta}_k - \boldsymbol{\beta}_0||_2^2/(2\lambda_1)$ and $R_0(\boldsymbol{\beta}_0) = ||\boldsymbol{\beta}_0||_2^2/(2\lambda_0)$ is consistent with the commonly utilized hierarchical normal prior model with, *a priori*, correlations $\mathrm{Cor}(\beta_{k,j}, \beta_{k',j}) = \lambda_0/(\lambda_0 + \lambda_1) > 0$ for studies $k' \neq k$ when

the latent vector $\boldsymbol{\beta}_0$ is integrated out. This regularization implies positive and symmetric borrowing of information for all pairs $k \neq k'$ of studies, and may not be appropriate for groups of studies with different patient populations.

For the latent mean parameter $\boldsymbol{\beta}_0$ we use the elastic-net penalty [54],

$$R_0(\boldsymbol{\beta}_0) = \lambda_0 ||\boldsymbol{\beta}_0||_1 + \lambda_1 ||\boldsymbol{\beta}_0||_2^2,$$

$\lambda_0, \lambda_1 \geq 0$, with LASSO and ridge penalty as special cases, when $\lambda_1 = 0$ and $\lambda_0 = 0$ respectively.

To account for differences and similarities of the available studies, we use

$$R_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} ||\boldsymbol{\beta}_{1:K,j} - \beta_{0,j}\mathbf{1}||_{\boldsymbol{\Sigma}}^a \tag{3}$$

in (1), where $\mathbf{1}$ is a K-dimensional vector with one on each component, $\boldsymbol{\beta}_{1:K,j} = (\beta_{1,j}, \cdots, \beta_{K,j})'$ refers to covariate $j = 1, \cdots, p$ in each study, and $||\boldsymbol{x}||_{\boldsymbol{\Sigma}} = \sqrt{\boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$. The symmetric matrix $\boldsymbol{\Sigma}$ is positive-semidefinite and enables differential borrowing of information across studies.

For $a = 2$, the minimizer of (1) is equivalent to the posterior mode when, *a priori*, the coefficients $\boldsymbol{\beta}_{1:K,j}, j = 1, \cdots, p$ across studies are modeled as multivariate normal with mean $\beta_{0,j}\mathbf{1}$ and covariance matrix $\boldsymbol{\Sigma}$. In this case $\Sigma_{k,k'} = 0$ implies that $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_{k'}$ are, *a priori* and conditionally on $\boldsymbol{\beta}_0$ independent. Whereas a large covariance $\Sigma_{k,k'} > 0$ indicates similarities between $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_{k'}$.

For $a = 1$ the penalty $R_1(\boldsymbol{\beta})$ in (3) becomes the sum of Mahalanobis distance of $\boldsymbol{\beta}_{1:K,j}$ from the mean $\beta_{0,j}\mathbf{1}$ with covariance matrix $\boldsymbol{\Sigma}$. With $\boldsymbol{\Sigma} \propto \boldsymbol{I}$ this penalty reduces to the group LASSO [10, 29] with one group for each covariate $j = 1, \cdots, J$.

The regularization parameters λ_0, λ_1, a, and $\boldsymbol{\Sigma}$ determine (i) the sparsity of $\widehat{\boldsymbol{\beta}}_0$ (the number of components $\widehat{\beta}_{0,j} = 0$) and (ii) the similarity of the estimates $\widehat{\beta}_{1,j}, \cdots, \widehat{\beta}_{K,j}$ across

studies, including the number of identical study-specific estimates $\widehat{\beta}_{k,j} = \widehat{\beta}_{k',j}$.

When $a \geq 1$, $\boldsymbol{\Sigma}$ is positive-definite and $\lambda_0 > 0$ or $\lambda_1 > 0$, the regularized log-partial-likelihood (1) is concave. If we fix $\lambda_1 \geq 0$, $a \geq 1$ and the positive-definite matrix $\boldsymbol{\Sigma}$, then the number of components $\widehat{\beta}_{0,j}$ equal to 0 increases with λ_0. For instance, consider $a > 1$ in (1), $\lambda_1 = 0$, and $g(\boldsymbol{\beta}_0) = \max_{\boldsymbol{\beta}_1,\cdots,\boldsymbol{\beta}_K} \sum_{k=1}^{K} l\left(\boldsymbol{\beta}_k, \mathcal{D}_k\right) - R_1\left(\boldsymbol{\beta}\right)$. The concave map $g(\boldsymbol{\beta}_0) - \lambda_0 ||\boldsymbol{\beta}_0||_1$ bounds the regularized log-partial-likelihood. If we choose λ_0 larger than $\max_{1\leq j\leq p} |\partial g(\boldsymbol{\beta}_0)/\partial \beta_{0,j}|$ at $\boldsymbol{\beta}_0 = \mathbf{0}$, then (1) is maximized at $\widehat{\boldsymbol{\beta}_0} = \mathbf{0}$. In contrast, for values λ_0 below this maximum some of the estimates $\widehat{\beta}_{0,j}$ will be different from zero.

For $\lambda_0, \lambda_1 \geq 0$, and $a = 1$, the choice of $\boldsymbol{\Sigma}$ can lead to identical study specific estimates $\widehat{\beta}_{1,j} = \cdots = \widehat{\beta}_{K,j}$. We provide an example with $\lambda_0 = 0$ and $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$. Let $\boldsymbol{z}_k = \boldsymbol{\beta}_k - \boldsymbol{\beta}_0, k = 1, \cdots, K$, and define $h(\boldsymbol{z}) = \max_{\boldsymbol{\beta}_0} \sum_{k=1}^{K} l\left(\boldsymbol{z}_k + \boldsymbol{\beta}_0, \mathcal{D}_k\right) - R_0\left(\boldsymbol{\beta}_0\right)$. The map $h(\boldsymbol{z}) - \sum_{1\leq j\leq p} ||(z_{1,j}, \ldots, z_{K,j})||_{\boldsymbol{\Sigma}}^{a}$ bounds the re-parametrized regularized log-partial-likelihood ($l_R : [\boldsymbol{\beta}_0, \boldsymbol{z}_1, \cdots, \boldsymbol{z}_K] \to \mathbb{R}$). If we specify $1/\sigma > \max_{j,k} |\partial h(\boldsymbol{z})/\partial z_{k,j}|$ at $\boldsymbol{z} = \mathbf{0}$, then the concave function (1) is maximized at $\widehat{\boldsymbol{\beta}_1} = \ldots = \widehat{\boldsymbol{\beta}_K} = \widehat{\boldsymbol{\beta}_0}$. More generally, if we don't assume a diagonal $\boldsymbol{\Sigma}$ and indicate with σ^2 the largest eigenvalue of $\boldsymbol{\Sigma}$, then the equalities $\widehat{\boldsymbol{\beta}_k} = \widehat{\boldsymbol{\beta}_0}$ hold when $1/\sigma > \max_{j,k} |\partial h(\boldsymbol{z})/\partial z_{k,j}|$ at $\boldsymbol{z} = \mathbf{0}$.

3 Parameter estimation

We use an alternating direction method of multipliers algorithm [7] to estimate $\boldsymbol{\beta}$, see [7] for an introduction to this algorithm. We first formulate the optimization of (1) with respect to $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_K)$ as a constrained convex minization problem

$$\min_{(\boldsymbol{\beta},\boldsymbol{z})} \left\{ \sum_k -l(\boldsymbol{\beta}_k, \mathcal{D}_k) + R_0(\boldsymbol{\beta}_0) + R_1(\boldsymbol{z}) \right\},$$

where $\boldsymbol{z} = (\boldsymbol{z}_0, \cdots, \boldsymbol{z}_K)', \boldsymbol{z}_k \in \mathbb{R}^p$, subjected to the affine constraints $\boldsymbol{\beta}_k = \boldsymbol{z}_k, k = 0, \ldots, K$. We then introduce for this minimization problem the scaled augmented Lagrangian

$$L_\rho(\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{u}) = \sum_{k=1}^{K} -l(\boldsymbol{\beta}_k, \mathcal{D}_k) + R_0(\boldsymbol{\beta}_0) + R_1(\boldsymbol{z}) + \sum_{k=0}^{K} \frac{\rho}{2} ||\boldsymbol{\beta}_k - \boldsymbol{z}_k + \boldsymbol{u}_k||_2^2, \tag{4}$$

where $\rho > 0$, with augmented $\boldsymbol{u} = (\boldsymbol{u}_0, \cdots, \boldsymbol{u}_K), \boldsymbol{u}_k \in \mathbb{R}^p$. For a fixed $\rho > 0$, the algorithm that we describe converges to a solution $\boldsymbol{\beta} - \boldsymbol{z} = \boldsymbol{u} = \boldsymbol{0}$ that maximizes (1). The algorithm minimizes (4) iteratively (i) with respect to $\boldsymbol{\beta}$, and (ii) with respect to $\boldsymbol{z}$, and (iii) then it updates $\boldsymbol{u}$ to $\boldsymbol{u} \leftarrow \boldsymbol{u} + \boldsymbol{\beta} - \boldsymbol{z}$, while keeping at each of the three steps the remaining two parameters fixed. At each iteration of the algorithm the minimization of (4) with respect to $\boldsymbol{\beta}$ (*step i*) can be carried out independently for each component $\boldsymbol{\beta}_k$, $k = 0, \ldots, K$, and the minimization with respect to $\boldsymbol{z}$ (*step ii*) can be carried out independently by covariates $j = 1, \cdots, p$.

The algorithm starts with an initial estimate of $\boldsymbol{\beta}$ (we use $\boldsymbol{0}$ or preliminary estimates of $\boldsymbol{\beta}_k, k = 0, \cdots, K$), $\boldsymbol{\beta} = \boldsymbol{z}$ and $\boldsymbol{u} = \boldsymbol{0}$. At each iteration, in *step i* the algorithm minimizes (4) $\boldsymbol{\beta}$, keeping $\boldsymbol{z}$ and $\boldsymbol{u}$ fixed, by setting $\boldsymbol{\beta}_0 = \dfrac{S\big(\rho(\boldsymbol{z}_0 - \boldsymbol{u}_0), \lambda_0\big)}{\rho + 2\lambda_1}$ where $S(\boldsymbol{x}, \lambda)$ is the coordinate-wise soft-thresholding function $s(x_j, \lambda) = (1 - \lambda/|x_j|)_+ x_j$ [45], and

$$\boldsymbol{\beta}_k = \arg\min_{\boldsymbol{b}} \Big(- l(\boldsymbol{b}, \mathcal{D}_k) + \rho ||\boldsymbol{b} - \boldsymbol{z}_k + \boldsymbol{u}_k||_2^2/2 \Big)$$

for the remaining $k = 1, \cdots, K$. We used a low-memory quasi-Newton algorithm [8] for the latter minimization $k > 0$.

In *step ii*, the algorithm minimizes (4) with respect to $\boldsymbol{z}$ keeping $\boldsymbol{\beta}$ and $\boldsymbol{u}$ fixed. This is done independently for each covariate $1 \leq j \leq p$, because $R_1(\cdot)$ and the l_2^2-norm in (4) can be factors into the sum of p terms each involving only the j-th row of $\boldsymbol{z} = (\boldsymbol{z}_0, \cdots, \boldsymbol{z}_K)$ and the j-th row of $\boldsymbol{\beta} + \boldsymbol{u}$. For example, when $a = 2$ in (3), $\boldsymbol{z}' = \Big(\boldsymbol{I} + 2\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H}/\rho\Big)^{-1} (\boldsymbol{\beta} + \boldsymbol{u})'$, where the K by $K + 1$ matrix $\boldsymbol{H} = [-\boldsymbol{1}, \boldsymbol{I}]$ is the concatenation of $-\boldsymbol{1}$ with k-dimensional

identity matrix $\boldsymbol{I}$. This, computation is implemented by first computing the matrix $K+1$ by $K+1$ matrix $\left(\boldsymbol{I}+2\boldsymbol{H}'\boldsymbol{\Sigma}^{-1}\boldsymbol{H}/\rho\right)^{-1}$, and then multiplying it with each column $j=1,\cdots,p$ of $(\boldsymbol{\beta}+\boldsymbol{u})'$.

Lastly, in *step iii*, $\boldsymbol{u}$ from the last iteration is updated to $\boldsymbol{u}+\boldsymbol{\beta}-\boldsymbol{z}$. We iterate this three steps until the l_2^2-norm of both $\boldsymbol{z}-\boldsymbol{\beta}$ and the difference between $\boldsymbol{z}$ from two successive iterations becomes smaller than a pre-specified treshold $\epsilon>0$ [7].

4 Simulation Study

We consider a total of 18 studies. Either $2, 5, 10$ or 15 of these 18 studies are used to estimate the model (1). The remaining $16, 13, 8$ or 3 studies are used for out-of study evaluations. For each study $k=1,\cdots,18$, we drew the sample size n_k of the study from a uniform distribution $n_k \sim Unif(100,\cdots,500)$, and then generated the covariates $x_{k,i}\in\mathbb{R}^{500}$ of observations $i=1,\cdots,n_k$ from a normal distribution $\boldsymbol{x}_{k,i}\sim N_{500}(\boldsymbol{0},\boldsymbol{V})$ with covariance $V_{j,j'}=0.3^{|j-j'|}$ between variables j and j'.

We then generated 100 times the parameters $\boldsymbol{\beta}\in\mathbb{R}^{500\times 19}$ and a collection of 18 studies $\mathcal{D}=(\mathcal{D}_k)_{k=1}^{18}$. In each of these 100 simulations we first generated the vector $\boldsymbol{\beta}_0\in\mathbb{R}^{500}$ from a two-component mixture distribution with (i) a point mass at zero and (ii) a normal distribution with mean zero and variance 0.1. The proportion of zeros of this mixture distribution equals $p_0=0.9$ or 0. We then generated independently $p=500$ vectors $(\epsilon_{1,j},\ldots,\epsilon_{K,j})\sim N_K(\boldsymbol{0},\boldsymbol{\Sigma})$ and set $\beta_{k,j}=\beta_{0,j}+\epsilon_{k,j}$ for each covariate $j=1,\ldots,p$ and study $k=1,\ldots,K$. We consider three matrices $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2,\boldsymbol{\Sigma}_3$ (see Figure 1) with 3, 2 or a single cluster of studies.

Survival times where generated from proportional hazard models with baseline survival functions $\widehat{S}_k(\cdot)$, regression coefficients $\boldsymbol{\beta}_k$, and censoring survival functions $\widehat{S}_{C,k}(\cdot)$. Here $\widehat{S}_k(\cdot)$ and $\widehat{S}_{C,k}(\cdot)$ have been estimated from the ovarian cancer datasets that we discuss in Section 5. For each study k we also generated an additional 1,000 observations, that were

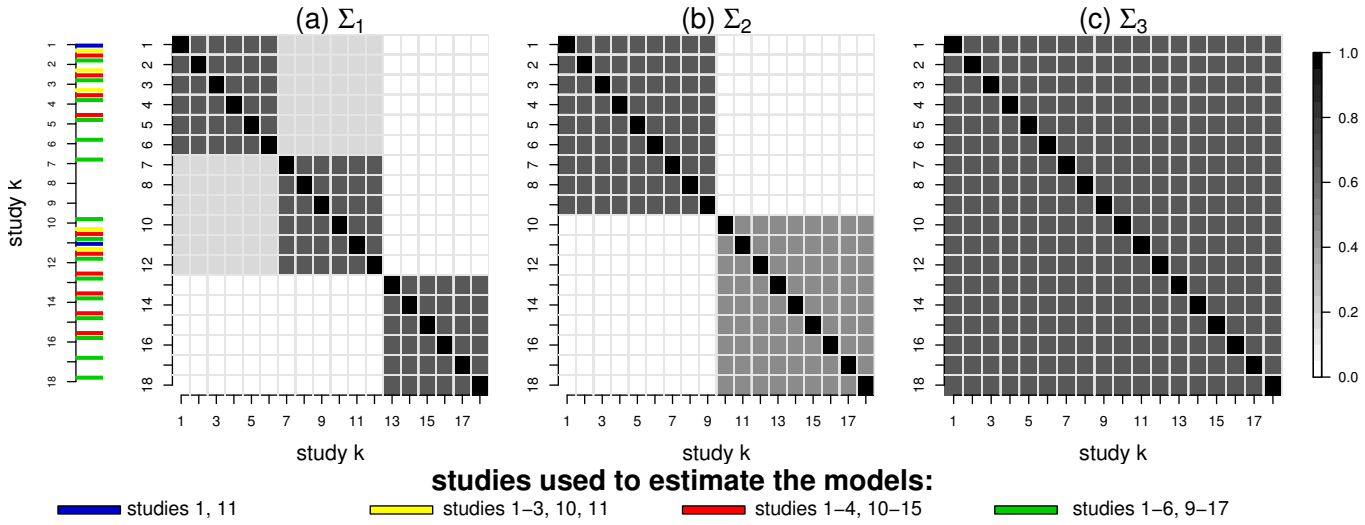

Figure 1: Similarity matrixes $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3 \in \mathbb{R}^{18\times 18}$ used to simulate the datasets. We simulated collections of 18 datasets $\mathcal{D} = \{D_k\}_{k=1}^{18}$ with similarity matrix $\boldsymbol{\Sigma}$ for $\boldsymbol{\beta}$. Studies indicated in blue (2 studies), yellow (5 studies), red (10 studies) or green (15 studies) are used to fit models with $K = 2, 5, 10$ or 15 studies.

not used to fit regression models, but were used to evaluate predictions.

4.1 Estimation of Σ and selection of (λ_1, λ_0)

We use initial estimates $\widehat{\boldsymbol{\beta}_k}$ obtained from K independent ridge regression models to estimate $\boldsymbol{\Sigma}$. The procedure leverage the Bayesian interpretation (2) of the regularized likelihood (1). As formalized in (2), with $R_1(\boldsymbol{\beta}) = \sum_{j=1}^{p} ||\boldsymbol{\beta}_{1:K,j} - \beta_{0,j}\mathbf{1}||_{\boldsymbol{\Sigma}}^{a}$, we can interpret $(\beta_{j,1}, \cdots, \beta_{j,K}), j = 1, \cdots, p$, as p independent vectors each with covariance matrix $\boldsymbol{\Sigma}$. If the $\boldsymbol{\beta}_k, k = 1, \cdots, K$, were known we could straightforwardly estimate $\boldsymbol{\Sigma}$. For instance with $a = 2$, the parameters $(\beta_{j,1}, \cdots, \beta_{j,K}), j = 1, \cdots, p$ can be interpreted as independent multivariate normal vectors with mean zero and covariance matrix $\boldsymbol{\Sigma}$. The joint normal distribution implies that $E[\boldsymbol{\beta}_k | \{\boldsymbol{\beta}_{k'}\}_{0<k'\leq K, k'\neq k}] = \sum_{0<k'\leq K, k'\neq k} \alpha_{k,k'}\boldsymbol{\beta}_{k'}$ where the weight vector $\boldsymbol{\alpha}_k = (\alpha_{k,k'})_{0<k'\leq K, k'\neq k}$ is as function of $\boldsymbol{\Sigma}$ [18] for each $k = 1, \cdots, K$. Therefore the conditional expectation of $\boldsymbol{X}_k\boldsymbol{\beta}_k$, given $\{\boldsymbol{\beta}_{k'}\}_{0<k'\leq K, k'\neq k}$ is $\sum_{0<k'\leq K, k'\neq k} \alpha_{k,k'}(\boldsymbol{X}_k\boldsymbol{\beta}_{k'})$. After replacing $\boldsymbol{\beta}_{k'}$ with out initial estimates $\widehat{\boldsymbol{\beta}}_{k'}$, we estimate $\boldsymbol{\alpha}_k$ via a Cox model with $K - 1$

covariates $z_{k'} = \boldsymbol{X}_k \widehat{\boldsymbol{\beta}}_{k'}$ and regression coefficients $\boldsymbol{\alpha}_k$. We then use the empirical covariance matrix of $\boldsymbol{\beta}_k^\star = \sum_{0<k'\leq K, k'\neq k} \widehat{\alpha}_{k,k'} \widehat{\boldsymbol{\beta}_k}, k = 1, \cdots, K$ as an estimate of $\boldsymbol{\Sigma}$. Note that $\boldsymbol{\beta}_k^\star$ has a direct interpretation under the assumption that the independent vectors $(\beta_{1,j}, \ldots, \beta_{K,j})$ are in a linear subspace with dimension less than K.

Figure 2 shows averages across the 100 simulations of the estimated similarity matrix between studies for the largest model with $K = 15$ studies when $p_0 = 0$ (top row) and $p_0 = 0.9$ (bottom row). Figures 2 and 1 show that the algorithm of Section 4.1 on average recovers the similarity structure of the 15 studies.

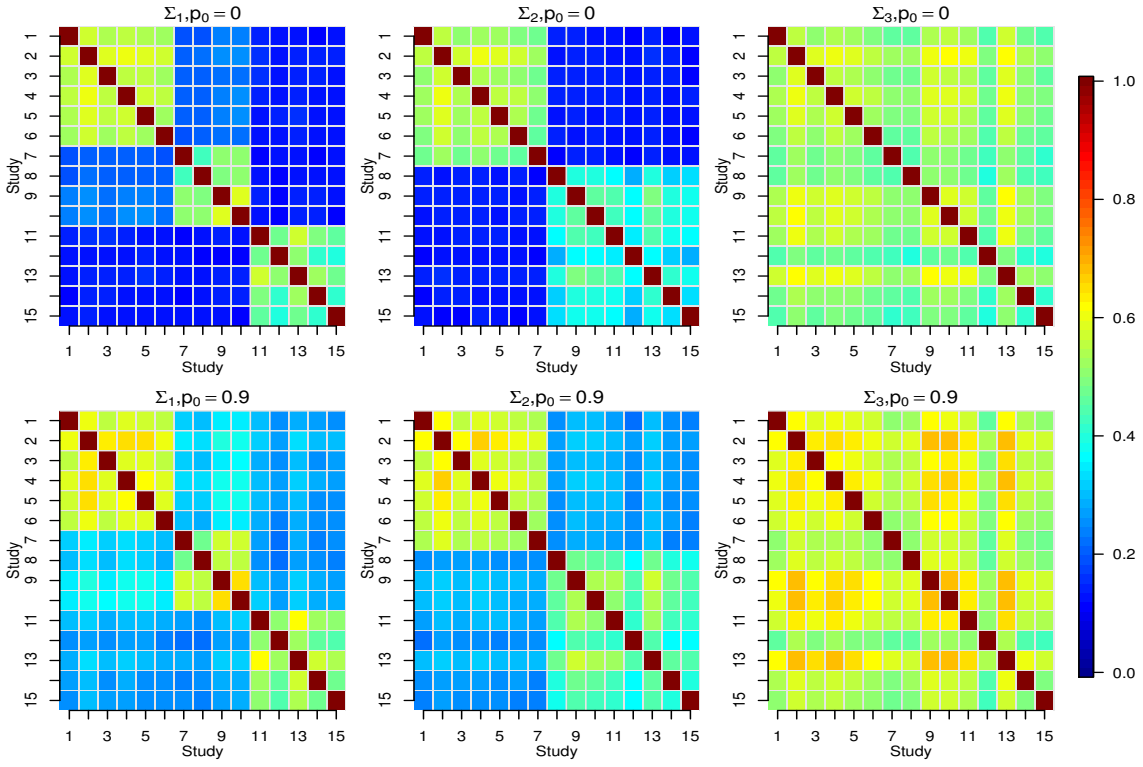

Figure 2: Average estimates across 100 simulations of the similarity matrix of the regression coefficients between the $K = 15$ studies.

To select the parameters λ_0 and/or λ_1, we use Monte-Carlo cross-validation (CV) [42]. We evaluate candidate parameter estimates $\widehat{\boldsymbol{\beta}}$ using $\mathcal{C}(\widehat{\boldsymbol{\beta}}) = \sum_k w_k \mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$, where the C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k) = \widehat{Pr}\left(\boldsymbol{x}_1' \widehat{\boldsymbol{\beta}}_k > \boldsymbol{x}_2' \widehat{\boldsymbol{\beta}}_k | Y_1 < Y_2\right)$ is the estimated concordance [21, 36, 49] between two survival times Y_1 and Y_2 with covariate vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in population k. The

weights $w_k \geq 0$ account for differences in study sample sizes, we used $w_k = 1/\sqrt{m_k}$. We first split the data randomly M-times into training (80%) and validation (20%) datasets. Next we define a grid of tuning parameters $\boldsymbol{\lambda} = (\lambda_1, \lambda_0)$. For each combination of tuning parameters $\boldsymbol{\lambda}$ of the grid, we estimate $\widehat{\boldsymbol{\beta}}^{(m)}$ based on the $m = 1, \cdots, M$ CV training datasets (which are identical across different grid-points) and use the validation datasets to obtain estimates of the study-specific C-statistics $\mathcal{C}_\lambda(\widehat{\boldsymbol{\beta}}_k^{(m)}, \mathcal{D}_k), m = 1, \cdots, M$ for $\widehat{\boldsymbol{\beta}}_k^{(m)}$ with $\boldsymbol{\lambda}$. We then average these M C-statistics and compute the overall estimate $\mathcal{C}_\lambda(\widehat{\boldsymbol{\beta}})$ for λ. Lastly, we select the λ-value with the highest average C-statistics $\mathcal{C}_\lambda(\widehat{\boldsymbol{\beta}})$.

4.2 Prediction Accuracy

Figure 3 shows, for each of the 18 studies box-plots of the estimated C-statistics [21, 36, 49] when either $2, 5, 10$ or 15 studies (1st to 4th column) were used to estimate the similarity matrix and the model. C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ for studies k that were utilized to estimate the model are highlighted inside the brown rectangles (estimated using the additional 1000 hold-out observations in study k), whereas C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ for studies k that were not used to estimate the model are shown on the right of the brown rectangles.

The three rows of Figure 3 correspond to scenarios with data generated using $\boldsymbol{\Sigma}_1$ (top row of Figure 3), $\boldsymbol{\Sigma}_2$, (2nd row), or $\boldsymbol{\Sigma}_3$ (bottom row) as illustrated in Figure 1. Red, green and blue box-plots on the top-row indicate the three clusters of studies under $\boldsymbol{\Sigma}_1$. Similarly, red and green box-plots in the 2-nd row indicate the two clusters of studies under Σ_2. Differences in the distribution of the C-statistics between studies within the same cluster are due to differences in the sample sizes n_k and covariate matrixes $\boldsymbol{X}_k$, which remain identical across the simulated datasets.

For $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ (1st row of Figure 3), with three clusters of studies, predictions show improvements when the number of studies used to train the regression models increases K. For $K = 2$ or 5, all studies used for estimation belong to the first two clusters (red and green

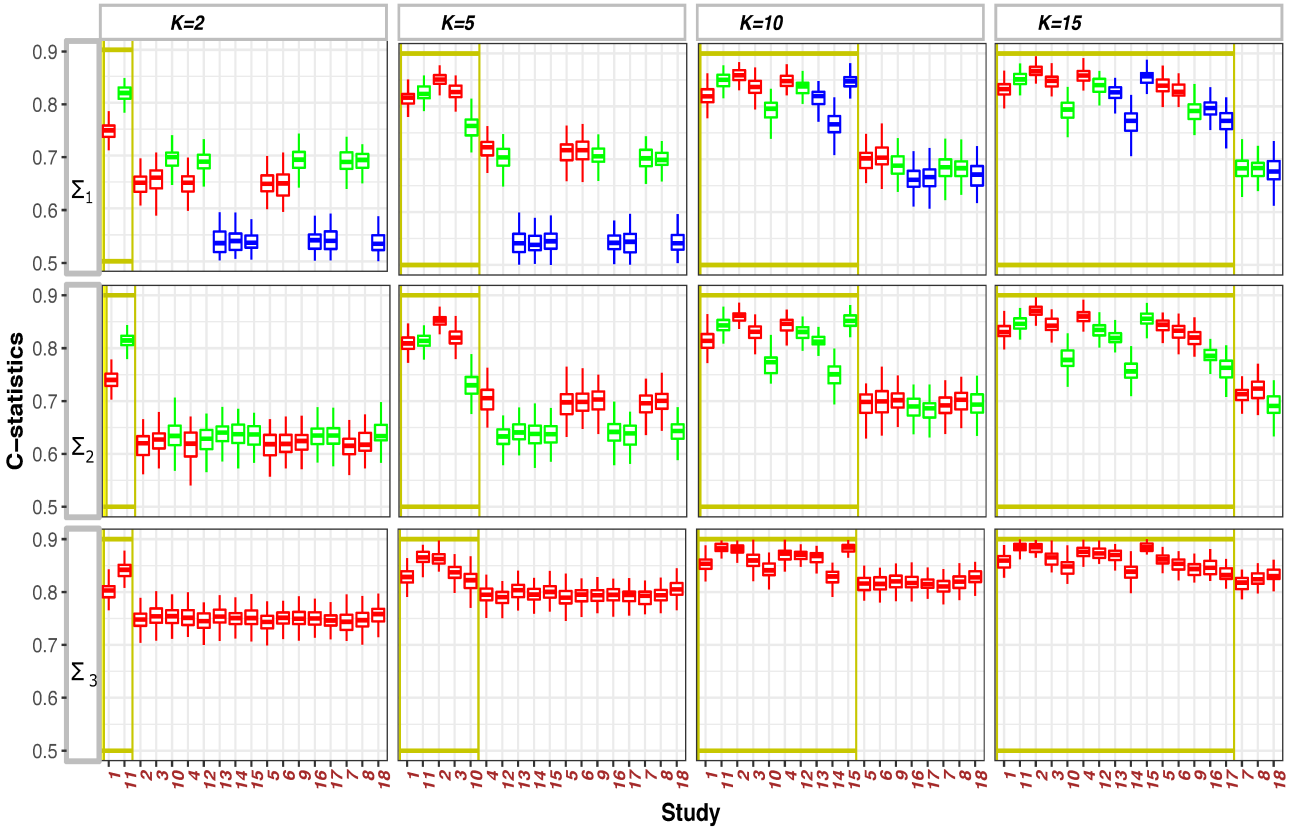

Figure 3: Predictions with the penalized regression model with $(a, \lambda_0) = (2, 0)$. We consider $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$ (see Figure 1), and $p_0 = 0$ across 100 simulations of a collection of 18 studies. Either $K = 2, 5, 10$ or 15 studies (studies inside the brown rectangles) were used for estimation/selection of $(\boldsymbol{\Sigma}, \lambda_1, \boldsymbol{\beta})$. The colors (red, green, blue) of the Box-plots indicate clusters of studies (3, 2, or 1 clusters when $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$ or $\boldsymbol{\Sigma}_3$).

box-plots). In these two cases, for each study $k = 13, \cdots 18$ in cluster 3 (blue box-plots) the inter-quartile range (IQR) of the C-statistics $\mathcal{C}(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ across simulations lies within the interval 0.52 to 0.55. Whereas for $K = 10$ (3rd column, studies 1-4 and 10-15 are use for estimation), studies from all three clusters have been used for training. In this case, the IQRs of $\mathcal{C}(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ across simulations for all three hold-out studies $k = 16, 17, 18$ in cluster 3 are within the interval 0.65 to 0.69. The last row of Figure 3 shows that, as expected, borrowing of information in the estimation of model parameters is most effective in the case of a single cluster of studies.

Next, we compared our estimates of $\boldsymbol{\beta}$ based on model (1), with $a = 2$ for $R_1(\cdot)$ and ridge penalty (HR-R, $\lambda_1 = 0$) or LASSO penalty (HR-L, $\lambda_2 = 0$) for $\boldsymbol{\beta}_0$, to Cox models trained separately on each study $\mathcal{D}_k$ with LASSO (single-study LASSO, SL) or ridge penalties (single-study ridge, SR) for $\boldsymbol{\beta}_k$. In addition we consider two models that combine all (2, 5, 10 or 15) studies into a single dataset and estimate a single Cox model (with regression parameters $\boldsymbol{\beta}_0$) using a LASSO (pooled LASSO, PL) or ridge (pooled ridge, PR) penalty for the coefficients $\boldsymbol{\beta}_0$. We also consider two meta-analysis approaches described in [51, 40] that combine study specific estimates $\widehat{\boldsymbol{\beta}}_k$ into a single vector $\widehat{\boldsymbol{\beta}}_0$ using either fixed effects (FE) or random-effects (RE) estimation.

Figures 4 and supplementary Figures A.1 and A.2 show the average C-statistics of each method when we used $K = 5$ or $K = 10$ studies for estimation. The pooled LASSO and ridge models (PL and PR) and the meta-analyses methods (FE and RE) estimate a single parameter $\boldsymbol{\beta}_0$, which was used to compute the C-statistics $C(\widehat{\boldsymbol{\beta}}_0, \mathcal{D}_k)$ for each study k. For the single-study SL and SR models we used the study-specific estimates $\widehat{\boldsymbol{\beta}}_k$ to compute $C(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ for in-study prediction (using the 1,000 validation observations). For prediction with SL and SR in studies k' not used for estimation, we used each estimate $\widehat{\boldsymbol{\beta}}_k$ of the $K = 5$ (or 10) training studies for predictions $\mathcal{C}(\widehat{\boldsymbol{\beta}}^k, \mathcal{D}_{k'})$ in all hold-out studies k'. For each hold-out studies k' we then averaged these $\mathcal{C}(\widehat{\boldsymbol{\beta}}^k, \mathcal{D}_{k'})$ over all $K = 5$ (or 10) training studies, i.e. Figure 4 reportes $\sum_k \mathcal{C}(\widehat{\boldsymbol{\beta}}^k, \mathcal{D}_{k'})/K$ for studied k'.

For studies k used to train the model, both HR-L and HR-R improve predictions $C(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ substantially compared to single-study estimates SL and SR. For instance, with $K = 5, p_0 = 0$ and unknown $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ (three clusters of studies), the average difference between $C(\widehat{\boldsymbol{\beta}}_k, \mathcal{D}_k)$ of HR-R and SR is between 0.07 and 0.16 for each of the five studies (0.62 to 0.75 for SR compared to 0.73 to 0.85 for HR-R). Similarly, meta-analytic and pooled estimates FE, RE and PL, SL improve predictions on the $K = 5$ datasets compared to single-study estimates, especially PR. But improvements are smaller than for HR-P and HR-L, with C-values on av-

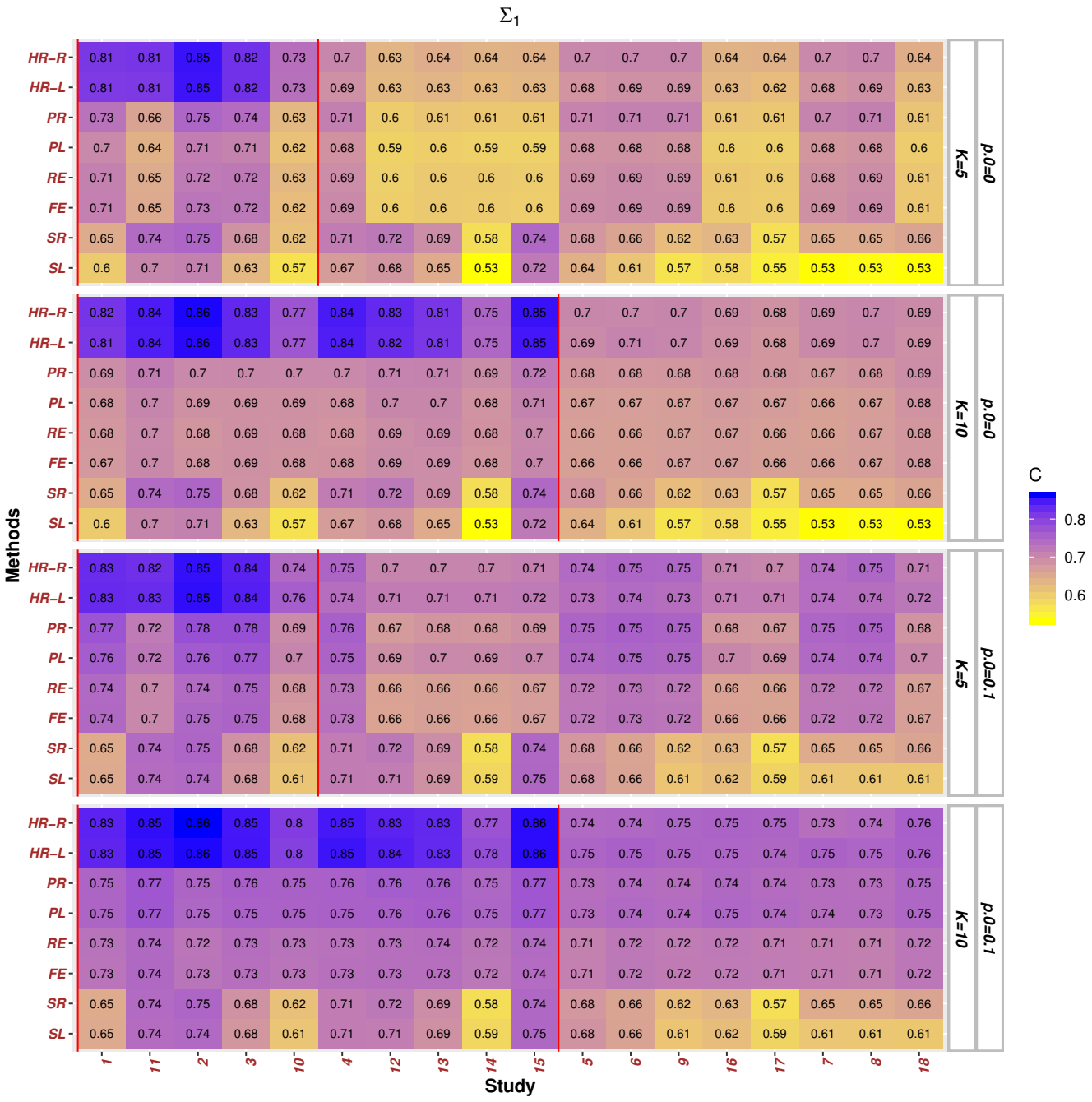

Figure 4: Average prediction across 100 simulations of a collection of 18 studies. Study specific effects $\boldsymbol{\beta}_k$ have been generated under $\boldsymbol{\Sigma}_1$ (see Figure 1). Either 5 or 10 of the 18 studies (studies in the left of the horizontal red bar) are used for the similarity matrix and covariate-effect estimation. See Figures A.1 and A.2 for results with similarity scenarios $\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3$.

erage (across simulations) between 0.08 to 0.15 below HR-R and HR-L models (for instance, 0.63 to 0.73 for PR compared to 0.73 to 0.85 for HR-R). When $K = 10$ studies are used to estimate the models, results are similar to the setting with $K = 5$ - meta-analytic, pooled and hierarchical estimates improve predictions over single-study estimates, with larger improvements for HR-R and HR-L estimates for in-study predictions.

In the case of a single cluster of studies ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_3$, supplementary Figure A.2), with strong similarity of the study-specific parameters $\boldsymbol{\beta}_k$, pooling of studies to estimate a single $\boldsymbol{\beta}_0$ is expected to be the most favorable prediction approach. Therefore, PL, PR, HR-R and HR-L predict survival substantially better than the FE, RE, SL and SR methods (supplementary Figure A.2). With $p_0 = 0$, in-study predictions based on HR-R and HR-L estimates are on average slightly better than for PR and PL estimates (difference of 0.01 to 0.04 for HR-R compared to PR with $K = 5$ studies, and 0.02 to 0.05 with $K = 10$). Whereas PR, PL, HR-R and HR-L have similar average C-statistics for holdout studies.

5 Survival prediction in ovarian cancer

We applied model (1) to predict survival in ovarian cancer using the *curatedOvarianData* repository, a curated collection of gene-expression datasets [20]. To evaluate prediction, we split the largest study in the database, the TCGA dataset [32] with 510 observations, 1,000 times randomly into a training dataset of $n_1 = 50, 75, \ldots,$ or 300 observations and a validation dataset with $510 - n_1$ observations. We predicted patient survival $Y_{1,i}$ in the TCGA holdout data by leveraging the hierarchical regularization model (1) using five additional datasets $k = 2, \cdots, K = 6$ (PMID-17290060[17], GSE51088[26], MTAB386[2], GSE13876[15] and GSE19829 [27]) with sample sizes ranging between $n_k = 42$ (GSE19829) and 157 (GSE13876) observations. In all the analyses we used the expression values of the $p = 3,030$ genes that are common in all six studies to predict patient survival.

To evaluate the hierarchical regularization method (1), we created different cross-study

heterogeneity scenarios that are motivated by documented inconsistencies across cancer datasets and by possible pre-processing errors [33, 39, 4, 1, 38, 41]. This is achieved by introducing in one (scenario 2: GSE13876[15]) or two studies (scenario 3: GSE13876[15] and GSE19829 [27]) a distortion of the expression values $x_{k,i,j}$ which become $10 - 3x_{k,i,j}, j = 1, \ldots, p$ for study $k = K$ (scenario two) or studies $k = K-1, K$ (scenario three). In scenario one we used the covariates $x_{k,i,j}$ of the six studies.

Similar to Section 4, we consider parameter estimates based on the $n_1 = 50, \cdots, 300$ TCGA training samples using (i) single study Cox models with LASSO (SL) or (ii) ridge (SR) regularization, pooled Cox regression models that combine the n_1 TCGA-observations and the remaining five studies (PMID-17290060, GSE51088, MTAB386, GSE13876 and GSE19829) into a single dataset with (iii) LASSO (PL) or (iv) ridge (PR) regularization, (v) fixed effects (FE) and (vi) random effects (RE) model meta-analyses models as described in [40, 3, 51], and (vii) the proposed hierarchical regularization model (1) with $\lambda_0 = 1, a = 2$ (HR-R).

Single-study cox-models with LASSO-penalty trained on the TCGA data with $n_1 = 50$ data points had low average C-values of 0.505 across the 1,000 generated training-validation samples, with minor improvements up 0.52 when $n_1 = 300$ observations are used for model training. Single study ridge regression models performed substantially better, with average C-statistics ranging between 0.53 for $n_1 = 50$ and 0.57 for $n_1 = 300$ observations. Improvements in risk predictions through integration of additional studies vary substantially across data-integration methods and scenarios. For scenario 1, FE and RE meta-analyses have both nearly constant and identical average C-statistics of 0.57 across all sample sizes n_1, while PL had an average C-statistics of 0.56 for $n_1 = 50$ with minor improvements up to 0.57 when $n_1 = 300$. Both, HR-R and PR have similar prediction accuracy across sample sizes n_1, with identical average C-statistics of 0.60 when $n_1 = 50$ and modest improvements up to 0.61 for both, HR-R and PR, when $n_1 = 300$.

Figure 5 shows, for scenarios two and three, average C-statistics for the TCGA validation samples. Different curves correspond to different prediction methods. The black curves show the average C-statistics (y-axis) across the 1,000 TCGA validation samples of size $510-n_1$ for Cox models trained on $n_1 = 50, \cdots, 300$ observations from the TCGA study (x-axis) using either SL (dotted curve) or SR (solid curve). The red curves show the average C-statistic for PR (solid curve) and PL (dotted curve) models, the green curves correspond to FE (dotted line) and RE (solid line) meta-analysis models, and the blue curve corresponds to the HR-R.

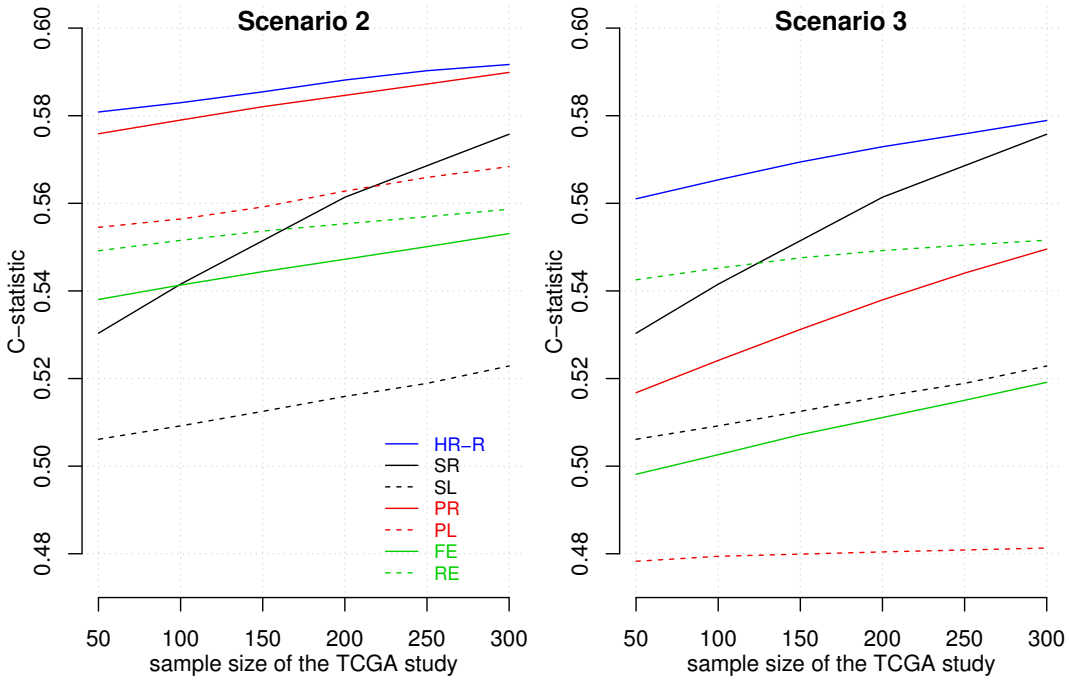

Figure 5: Average C-statistics for single-study SL and SR methods with $n_1 = 50, \cdots, 300$ TCGA training samples, and for data-integration methods (PL, PR, FE, RE, HR-R) use the n_1 TCGA training samples and training samples from five additional studies (PMID-17290060, GSE51088, MTAB386, GSE13876 and GSE19829).

In scenario two, the RE meta-analysis, which combines estimates from the $n_1 = 50$ TCGA data points with estimates from the remaining five studies, has the same average C-statistics as the single-study SR model trained on $n_1 = 240$ patients. For sample sizes $n_1 > 250$, pooled regression models PR have similar performances as RE models. The HR-R model trained on $n_1 = 50$ TCGA patients has an average C-statistics the is superior to those of

PR and FE procedures with $n_1 = 50, \cdots, 300$. As expected, with increased discrepancies in the relations between covariates and outcomes across studies (scenario three), performances of all data-integration methods decrease. HR-R models with $n_1 = 50$ TCGA patients have similar average C-values as single study SR modes with $n_1 \approx 200$ patients. PR, PL, FE and RE methods rely on the assumption that the regression parameters are similar across studies. With substantial departures from this assumption the hierarchical model HR-R shows, across all sample sizes $50 \leq n_1 \leq 300$, gains in average prediction accuracy compared to PR, PL, FE and RE.

6 Discussion

The analysis of relations between omics variables and time to event outcomes, and the use of individual profiles $\boldsymbol{x}_{k,i}$ for predictions, are particularly challenging when the sample size n_k is small. These analyses often include thousands of potential predictors. The use of multiple studies and pooling of information can improve prediction accuracy. Meta-analyses can be utilized when the relations of covariates and outcomes are homogeneous across studies. But recent work in oncology [40, 51, 47] showed that there can be clusters of studies with relevant discrepancies in their covariate-outcome relations due, for example, to differences in study designs, patient populations and treatments.

We combined two established concepts, regularization of regression models [25, 45, 46, 19] and metrics of similarity between datasets [22] that identify clusters of studies. We used these concepts to estimate study-specific regression parameters $\boldsymbol{\beta}_k$ and for predictions, both in $k = 1, \ldots, K$ contexts that are represented in our collection of datasets, for example K distinct geographic regions, and in other contexts ($k = K + 1$) by estimating the latent parameters $\boldsymbol{\beta}_0$.

The $K \times K$ similarity matrix $\boldsymbol{\Sigma}$ is used to regularize the likelihood function, and it tunes the degree of borrowing of information in the estimation of K study-specific regression

models. It shrinks the estimate of the k-th study-specific regression parameter $\boldsymbol{\beta}_k$ towards estimates $\boldsymbol{\beta}_{k'}$ of studies k' that are similar to study k (large $\Sigma_{k,k'}$). In contrast studies with low similarity ($\Sigma_{k,k'} \approx 0$) have little influence on the estimation of $\boldsymbol{\beta}_k$. In our analyses we verified that, if there are clusters of studies with similar predictors-outcome relations, then the introduced method improves the accuracy of predictions compared to alternative procedures, including single-study estimates, meta-analyses and pooling of all studies into a single data matrix.

Funding

Lorenzo Trippa was partially supported by the NSF grant 1810829.

References

[1] C. R. Acharya, D. S. Hsu, C. K. Anders, A. Anguiano, K. H. Salter, K. S. Walters, R. C. Redman, S. A. Tuchman, C. A. Moylan, S. Mukherjee, et al. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer. *Jama*, 299(13):1574–1587, 2008.

[2] S. Bentink, B. Haibe-Kains, T. Risch, J.-B. Fan, M. S. Hirsch, K. Holton, R. Rubio, C. April, J. Chen, E. Wickham-Garcia, et al. Angiogenic mrna and microrna gene expression signature predicts a novel subtype of serous ovarian cancer. *PloS one*, 7(2), 2012.

[3] C. Bernau, M. Riester, A.-L. Boulesteix, G. Parmigiani, C. Huttenhower, L. Waldron, and L. Trippa. Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112, 2014.

[4] H. Bonnefoi, A. Potti, M. Delorenzi, L. Mauriac, M. Campone, M. Tubiana-Hulin, T. Petit, P. Rouanet, J. Jassem, E. Blot, et al. Retracted: Validation of gene signatures

that predict the response of breast cancer to neoadjuvant chemotherapy: a substudy of the eortc 10994/big 00-01 clinical trial, 2007.

[5] H. M. Bøvelstad, S. Nygård, H. L. Størvold, M. Aldrin, Ø. Borgan, A. Frigessi, and O. C. Lingjærde. Predicting survival from microarray data a comparative study. *Bioinformatics*, 23(16):2080–2087, 2007.

[6] H. M. Bøvelstad, S. Nygård, and Ø. Borgan. Survival prediction from clinico-genomic models-a comparative study. *BMC bioinformatics*, 10(1):1, 2009.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[8] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[9] Y. Q. Chen and M.-C. Wang. Analysis of accelerated hazards models. *Journal of the American Statistical Association*, 95(450):608–618, 2000.

[10] X. Cheng, W. Lu, and M. Liu. Identification of homogeneous and heterogeneous variables in pooled cohort studies. *Biometrics*, 71:397–403, 2015.

[11] E. Conlon, B. Postier, B. Methe, K. Nevin, and D. Lovley. Hierarchical bayesian meta-analysis models for cross-platform microarray studies. *Journal of Applied Statistics*, 36 (10):1067–1085, 2009.

[12] E. M. Conlon, J. J. Song, and J. S. Liu. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC bioinformatics*, 7(1):1, 2006.

[13] E. M. Conlon, B. L. Postier, B. A. Methé, K. P. Nevin, and D. R. Lovley. A bayesian model for pooling gene expression studies that incorporates co-regulation information. *PloS one*, 7(12):e52137, 2012.

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

[15] A. P. Crijns, R. S. Fehrmann, S. de Jong, F. Gerbens, G. J. Meersma, H. G. Klip, H. Hollema, R. M. Hofstra, G. J. te Meerman, E. G. de Vries, et al. Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS medicine*, 6(2), 2009.

[16] R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

[17] H. K. Dressman, A. Berchuck, G. Chan, J. Zhai, A. Bild, R. Sayer, J. Cragun, J. Clarke, R. S. Whitaker, L. Li, et al. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *Journal of Clinical Oncology*, 25(5):517–525, 2007.

[18] M. L. Eaton. Multivariate statistics: a vector space approach. *JOHN WILEY & SONS*, 1983.

[19] W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

[20] B. F. Ganzfried, M. Riester, B. Haibe-Kains, T. Risch, S. Tyekucheva, I. Jazic, X. V. Wang, M. Ahmadifar, M. J. Birrer, G. Parmigiani, et al. curatedovariandata: clinically annotated data for the ovarian cancer transcriptome. *Database*, 2013, 2013.

[21] F. E. Harrell Jr, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression

modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2): 143–152, 1984.

[22] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

[23] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[24] L. V. Hedges and I. Olkin. *Statistical methods for meta-analysis*. Academic press, 2014.

[25] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[26] B. Y. Karlan, J. Dering, C. Walsh, S. Orsulic, J. Lester, L. A. Anderson, C. L. Ginther, M. Fejzo, and D. Slamon. Postn/tgfbi-associated stromal signature predicts poor prognosis in serous epithelial ovarian cancer. *Gynecologic oncology*, 132(2):334–342, 2014.

[27] P. A. Konstantinopoulos, D. Spentzos, B. Y. Karlan, T. Taniguchi, E. Fountzilas, N. Francoeur, D. A. Levine, and S. A. Cannistra. Gene expression profile of brcaness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *Journal of clinical oncology*, 28(22):3555, 2010.

[28] F. Liu, D. Dunson, and F. Zou. High-dimensional variable selection in meta-analysis for censored data. *Biometrics*, 67(2):504–512, 2011.

[29] M. Liu, W. Lu, V. Krogh, G. Hallmans, T. V. Clendenen, and A. Zeleniuch-Jacquotte. Estimation and selection of complex covariate effects in pooled nested case–control studies with heterogeneity. *Biostatistics*, 14(4):682–694, 2013.

[30] S. Ma, J. Huang, and X. Song. Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics*, 12(4):763–775, 2011.

[31] S. Ma, J. Huang, F. Wei, Y. Xie, and K. Fang. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in medicine*, 30(28): 3361–3371, 2011.

[32] C. G. A. R. Network et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609, 2011.

[33] N. D. of Health and H. Services. Findings of research misconduct. *NIH Guide Grants Contracts*, 16(021), 2015.

[34] M. Y. Park and T. Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69 (4):659–677, 2007.

[35] T. Park and G. Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[36] M. J. Pencina and R. B. D'Agostino. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*, 23(13):2109–2123, 2004.

[37] N. G. Polson, J. G. Scott, B. T. Willard, et al. Proximal algorithms in statistics and machine learning. *Statistical Science*, 30(4):559–581, 2015.

[38] A. Potti, A. Bild, H. K. Dressman, D. A. Lewis, J. R. Nevins, and T. L. Ortel. Gene-expression patterns predict phenotypes of immune-mediated thrombosis. *Blood*, 107(4): 1391–1396, 2006.

[39] A. Potti, H. K. Dressman, A. Bild, R. F. Riedel, G. Chan, R. Sayer, J. Cragun, H. Cottrill, M. J. Kelley, R. Petersen, et al. Genomic signatures to guide the use of chemotherapeutics. *Nature medicine*, 12(11):1294–1300, 2006.

[40] M. Riester, W. Wei, L. Waldron, A. C. Culhane, L. Trippa, E. Oliva, S.-h. Kim, F. Michor, C. Huttenhower, G. Parmigiani, et al. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *Journal of the National Cancer Institute*, page dju048, 2014.

[41] K. H. Salter, C. R. Acharya, K. S. Walters, R. Redman, A. Anguiano, K. S. Garman, C. K. Anders, S. Mukherjee, H. K. Dressman, W. T. Barry, et al. An integrated approach to the prediction of chemotherapeutic response in patients with breast cancer. *PLoS One*, 3(4):e1908–e1908, 2008.

[42] J. Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

[43] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software*, 39 (5):1, 2011.

[44] D. Sinha, J. G. Ibrahim, and M. Chen. A bayesian justification of cox's partial likelihood. *Biometrika*, 90(3):629–641, 2003. doi: 10.1093/biomet/90.3.629. URL +http://dx.doi.org/10.1093/biomet/90.3.629.

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

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

[47] L. Trippa, L. Waldron, C. Huttenhower, and G. Parmigiani. Bayesian nonparametric cross-study validation of prediction methods. *Ann. Appl. Stat.*, 9(1):402–428, 03 2015. doi: 10.1214/14-AOAS798.

[48] L. Trippa, L. Waldron, C. Huttenhower, G. Parmigiani, et al. Bayesian nonparametric cross-study validation of prediction methods. *The Annals of Applied Statistics*, 9(1): 402–428, 2015.

[49] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. Wei. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

[50] W. N. Van Wieringen, D. Kun, R. Hampel, and A.-L. Boulesteix. Survival prediction using gene expression data: a review and comparison. *Computational statistics & data analysis*, 53(5):1590–1603, 2009.

[51] L. Waldron, B. Haibe-Kains, A. C. Culhane, M. Riester, J. Ding, X. V. Wang, M. Ahmadifar, S. Tyekucheva, C. Bernau, T. Risch, et al. Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *Journal of the National Cancer Institute*, 106(5):dju049, 2014.

[52] H. Wang and C. Leng. A note on adaptive group lasso. *Computational Statistics & Data Analysis*, 52(12):5277 – 5286, 2008. ISSN 0167-9473. doi: http://dx.doi.org/10.1016/j.csda.2008.05.006. URL http://www.sciencedirect.com/science/article/pii/S0167947308002582.

[53] L. Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

[54] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

A Appendix

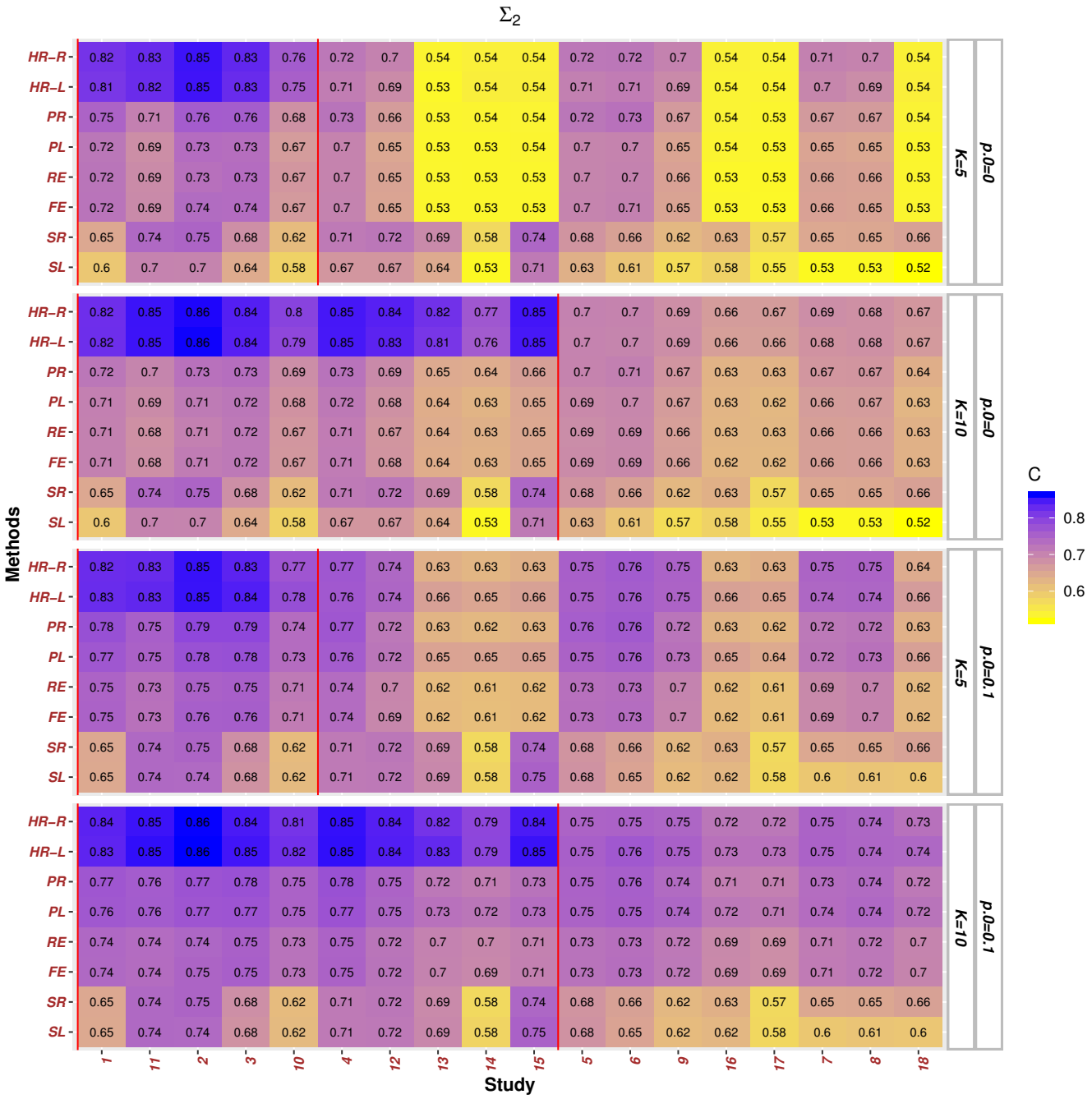

Figure A.1: Average prediction across 100 simulations of a collection of 18 studies. Study specific effects $\boldsymbol{\beta}_k$ have been generated under $\boldsymbol{\Sigma}_2$. Either 5 or 10 of the 18 studies (studies in the left of the horizontal red bar) are used for the similarity matrix and covariate-effect estimation.

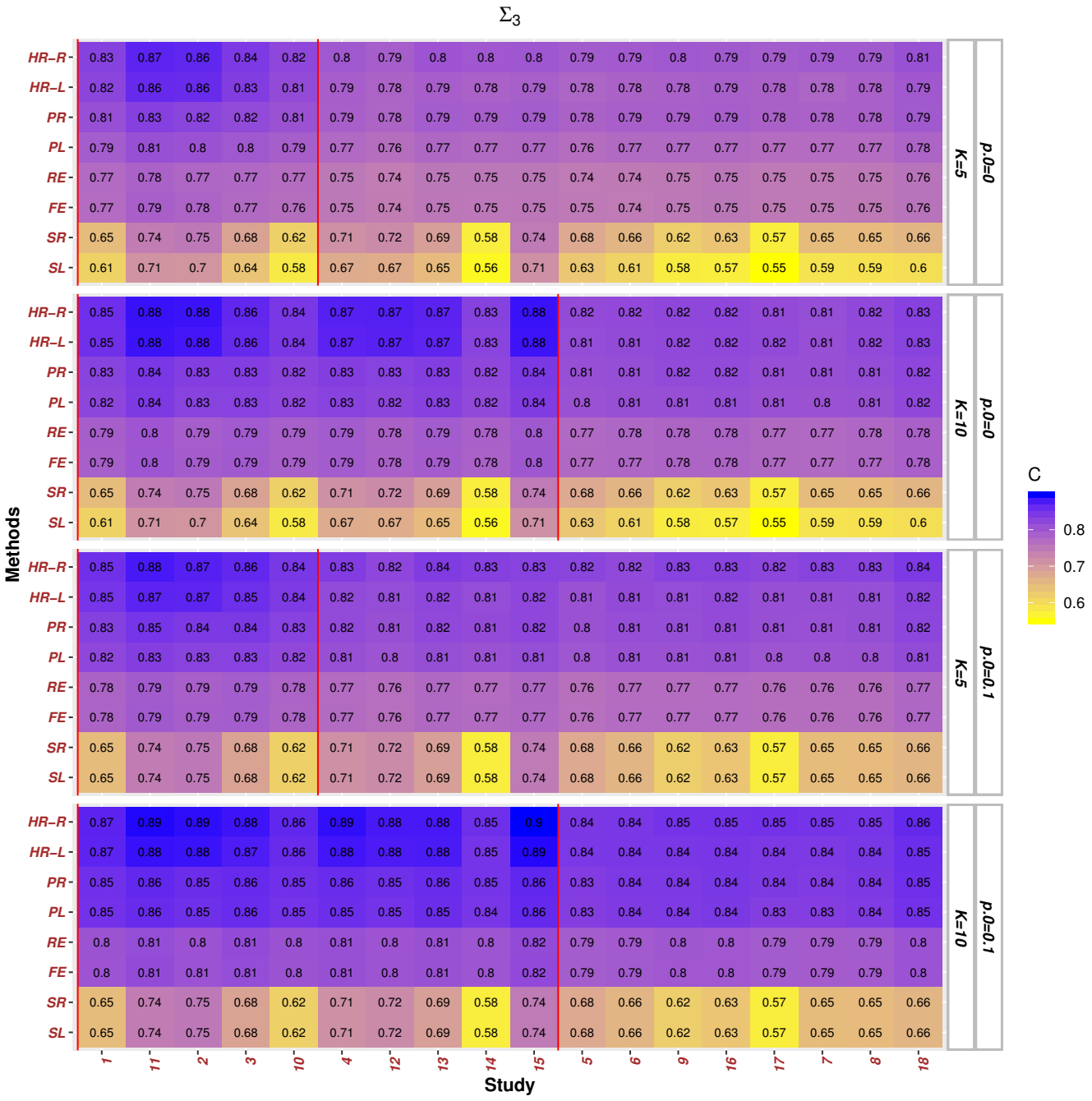

Figure A.2: Average prediction across 100 simulations of a collection of 18 studies. Study specific effects $\boldsymbol{\beta}_k$ have been generated under $\boldsymbol{\Sigma}_3$. Either 5 or 10 of the 18 studies (studies in the left of the horizontal red bar) are used for the similarity matrix and covariate-effect estimation.