Combining clinical and molecular data in regression prediction models: insights from a simulation study

Riccardo De Bin
Department of Mathematics
University of Oslo, Norway
(corresponding author – phone number: +47-22855859,
fax number: +47-22854349, email address: debin@math.uio.no)

Anne-Laure Boulesteix
Institute for Medical Information Processing, Biometry and Epidemiology
University of Munich, Germany

Axel Benner
Division of Biostatistics
German Cancer Research Centre of Heidelberg, Germany

Natalia Becker
Division of Biostatistics
German Cancer Research Centre of Heidelberg, Germany

Willi Sauerbrei
Institute of Medical Biometry and Statistics,
University of Freiburg (Germany)

Combining clinical and molecular data in regression prediction models: insights from a simulation study

Riccardo De Bin* Anne-Laure Boulesteix Axel Benner
Natalia Becker Willi Sauerbrei

Abstract

Data integration, i.e. the use of different sources of information for data analysis, is becoming one of the most important topics in modern statistics. Especially in, but not limited to, biomedical applications, a relevant issue is the combination of low-dimensional (e.g., clinical data) and high-dimensional (e.g., molecular data such as gene expressions) data sources in a prediction model. Not only the different characteristics of the data, but also the complex correlation structure within and between the two data sources, pose challenging issues. In this paper, we investigate these issues via simulations, providing some useful insight into strategies to combine low- and high-dimensional data in a regression prediction model. In particular, we focus on the effect of the correlation structure on the results, whilst accounting for the influence of our specific choices in the design of the simulation study.

1 Introduction

During the last decades the amount of molecular data collected has increased substantially. It is hoped that (multiple) omics profiles help to improve diagnosis, prognosis, therapy and more [42]. However, clinical data are generally available and it is an important question whether the combination of clinical and omics data in a joint model leads to better statistical properties (such as smaller prediction errors). Combining low-dimensional clinical and high-dimensional molecular sources of information in a prediction model,

*debin@math.uio.no - Department of Mathematics, University of Oslo, Norway

however, is not straightforward. Several issues arise due to their different nature: the former is characterized by few variables whose significance is usually well-validated in the biomedical literature; the latter by a large number of variables and a low signal to noise ratio. When simply pulled together in a procedure which involves variable selection, it is well known that the risk of "losing" informative clinical predictors among the large number of molecular variables is quite high [3, 4]. Correlation structures within and between the data sources worsen this situation.

Different strategies have been introduced in the literature to address the issue, nicely summarized by [4]. The main idea is to adapt statistical methods to fully exploit the clinical information notwithstanding the noise linked to the molecular data. When these strategies, combined with several statistical methods, have been applied to real data [see, e.g., 7, 16], no clear winner has emerged. This is not surprising, as comparably good statistical procedures are usually designed to perform best in different specific situations. Among the limitations of comparisons based on real data [see also 5], the impossibility of controlling the data structure is arguably the most severe: no clear recommendation can be provided to practitioners on which procedure performs best in specific situations. To reach this goal, analyses on simulated data must be performed. For example, Truntzer and colleagues [40] reinforced their comparison on the predictive ability of new and existing models combining mass spectrometry data and classical clinical variables on a binary outcome by conducting a simulation study. Using several approaches to derive models they conclude that the model based on only clinical variables was less efficient than the model based on mass spectrometry only. Nevertheless, they stated that "It is hard to decide which method is the best one".

From a slightly different prospective, Zhu and colleagues [46] investigated the integration of clinical and multiple omics data for prognostic assessments of fourteen tumour types. Their focus was on the diseases: they found that in seven of them the difference between the c-indices from the clinical model and the combined model was negligible (difference less than 0.01), while the combined model had larger c-index values in the other seven cancers. Canuel and colelagues [12], instead, focused on the translational research platforms for managing and exploring the integration of clinical and omics data, giving a critical review of seven possible solutions.

The goal of this paper is to perform an extensive simulation study contrasting several strategies for combining clinical and molecular information in a regression prediction model, focusing on the effect of different correlation structures. A simulation study in this context is by nature limited: even if the design tries to generate data as realistically as possible, it is extremely

difficult to fully reproduce the complexity of real data, and, even more so, to simulate the whole spectrum of situations which could be experienced in practice. Despite these limitations, or, better, by discussing the effect of these limitations, this study provides relevant insight into the problem of interest and into the use of a class of statistical methods, namely penalized regression approaches, when both clinical and molecular data are available.

For better understanding, transparency and complete reporting, we summarize key steps of the design and results in a two-part simulation profile (see Table 1). This profile is adapted from the REMARK profile which was introduced in the reporting recommendations for tumor marker prognostic studies, following the idea that a structured display helps to provide a better overview of the study and analyses. This therefore helps to avoid selective reporting and makes strengths and weaknesses more explicit [1].

The paper is structured as follows: after a short review of the methods used in the study (Section 2), the simulation design and the data generating process are discussed in Section 3. Section 4 contains some remarks on the correlation and its effect in the model estimation. The numerical results and the related comments are provided in Section 5, while an illustrative example on real data is shown in Section 6. Finally, some remarks and recommendations are presented in Section 7. Further results and the R code to reproduce the analyses are given in the Supplementary Material.

2 Methods

2.1 Strategies to combine clinical and molecular variables

As mentioned above, some care is necessary to combine clinical and molecular data in a prediction model. Here we briefly outline four strategies as previously described by [4] and [16].

2.1.1 Strategy 1: "naive"

The most straightforward and "naive" way to combine clinical and omics variables is to simply treat them equally, i.e., to fully ignore their different nature when estimating their coefficients. This strategy, indexed as '1' from now on, is very easy to implement, but has the major pitfall that clinical information is potentially masked by the high number of omics variables and might not be fully exploited by the model [3, 4]. Prior knowledge most often tells us that clinical variables are on average more informative than

a) Design			
Question	Comparing the prediction ability of strategies which combine clinical and molecular variables (C and M variables)		
Combinations	Seven strategies to combine C and M variables, five methods to construct a prediction model, preliminary screening (yes/no), giving 70 strategy/method/screening combinations		
Strategies	Naive, Clinical offset, Favoring, Dimension Reduction. All with/without clinical variable selection, apart from Naive		
Methods	Boosting, Lasso, Ridge, Elastic net, SCAD		
Screening	Sure Independent Screening (SIS). We tried with Iterative Sure Independent Screening (ISIS), but it never converged. Will be ignored		
Variables	15 clinical variables (5 with and 10 without effect) 10000 molecular variables in 50 independent blocks, 28 variables with effect (see Table 2)		
Correlation	Structured within blocks of C and M variables and between the blocks (no [0], moderate [0.5], strong [0.8] correlation) Nine settings (see Table 3), 3 settings presented in detail, others in the Supplementary Material.		
Sample Size	500 (100 and 1000 in the Supplementary Material)		
Outcome	Mean Square Prediction Error (MSPE), Sensitivity (true positive rate) and Specificity (true negative rate).		
b) Results			
Setting	MSPE	Sens/spec	Remarks
B1: set 1, no correlation, no pre-screening	Tab 5 for SCAD (Fig 1a) for favor.2 (Fig 1b) (ridge excluded)	For SCAD clin. dat. (Fig 3) mol. dat. (Fig 4) for favor.2 (Fig 5)	SCAD/favor.2 best performance MSPE
B2: set 2, high correlation, no pre-screening	Tab 6 for boosting (Fig 1c) for dim.red.1 (Fig 1d) (ridge excluded)	For boosting clin. dat. (Fig 3) mol. dat. (Fig 4) for favor.2 (Fig 5)	Boosting/dim.red.1 best performance MSPE
B3: set 3, mod. correlation, no pre-screening	Tab 7 for boosting (Fig 1e) for dim.red.1 (Fig 1f) (ridge excluded)	For boosting clin. dat. (Fig 3) mol. dat. (Fig 4) for favor.2 (Fig 5)	Boosting/dim.red.1 best performance MSPE
B4: effect of pre-screening	Fig 6		Only beneficial for ridge regression
B5: set 3 to 8	Suppl. Material	Suppl. Material	

Table 1: Simulation profile.

omics variables, but this prior knowledge is not exploited. This may lead to a model with a sub-optimal prediction accuracy, especially if the clinical variables have strong effects.

2.1.2 Strategy 2: "clinical offset"

To prevent the "masking" of clinical variables, it is possible to force them into the model. This can be done by first fitting a model to the clinical variables only, and then using the resulting linear predictor (hereafter denoted as *clinical linear predictor*) as an offset (i.e. as a variable with coefficient fixed to 1) in a prediction model fitted to the molecular variables. This strategy is denoted as *clinical offset* strategy [16] and indexed as '2' from now on. Compared to the naive strategy, which ignores the different nature of clinical and omics variables, this strategy can be seen as the other extreme. Indeed, the estimates of the coefficients of the clinical variables are not allowed to change when included in the combined model; therefore the interplay between clinical and molecular variables may not be fully taken into consideration—possibly leading to sub-optimal prediction accuracy.

This strategy may make sense when the effects of the clinical variables are well validated in the literature. However, it may happen that other clinical variables are provided, which may not all have a relevant effect. A natural question is whether we should derive the clinical offset using all the clinical variables or first apply a variable selection procedure (in this work, backward elimination with AIC stopping criterion) to focus on the relevant clinical variables. In this paper we assess both strategies, referring to them as "2.1" (without variable selection) and "2.2" (with variable selection).

2.1.3 Strategy 3: "favoring"

A different strategy, which can be seen as a compromise between the naive strategy and the clinical offset strategy, consists of fitting a model using both clinical and molecular variables simultaneously, but giving more "weight" to the former. This strategy is termed *favoring strategy* [4, 16] and indexed as '3' from now on. Since in our analysis all the regression techniques are based on penalized regression in a broad sense, we can define the penalty term in such a way that molecular variables are more strongly penalized. More precisely, here we impose a penalty (obtained by cross-validation) only to the molecular variables, leaving the clinical ones unpenalized. Similarly to the clinical offset strategy, we consider a version of this strategy in which all clinical variables are used ("3.1") and one in which we only consider the clinical variables selected by a variable selection procedure ("3.2").

2.1.4 Strategy 4: "dimension reduction"

Another possibility to tackle the difference in dimensionality between clinical and molecular variable spaces is to summarize the molecular information into a "score" (i.e., a linear combination of selected variables), that is later included in a regression model along with the clinical variables. In this way, the number of variables related to the molecular data is comparable to that of the clinical variables. Similarly to strategies 2.1/2.2 and 3.1/3.2, we either consider all clinical variables, i.e. we add the molecular score to the full clinical model ("4.1"), or a selection of them, i.e. we add the molecular score to the reduced model ("4.2").

2.2 Methods to derive a prediction model

Among the large number of methods developed to handle high-dimensional data, those based on penalized regression are among the most used. In this section we quickly review the five approaches considered in this paper. They have been selected due to their popularity and the presence of user-friendly R [32] packages. Note that we strongly rely on the functions contained in these packages to choose the values of the tuning parameters. In order to be as fair as possible, we try not to modify the default procedures and settings, as would probably be done by some analysts. An important exception is made as far as the choice of the internal validation procedure is concerned. For the purpose of better comparability, we apply the 10-fold cross-validation technique for parameter tuning, thus avoiding that the accuracy of the cross-validation technique influences the method performance. Further details on the methods' implementation considered in our paper are provided in the respective sections.

2.2.1 Boosting

The boosting approach is based on the idea of repeatedly applying a weak estimator to a modification of the data to minimize a loss function. Originally developed in the machine learning community as a classifier, it has been translated to cope with numerous statistical problems [see, e.g., 35], including, relevantly for this paper, linear regression. In the case of linear regression, the commonly used loss function is the quadratic loss and the weak estimator, in our case a penalized version of the least square estimator, is repeatedly fitted to the residuals. The estimates of the regression coefficients are updated in a forward stepwise manner, until a pre-specified number of steps is reached. Since we deal with high-dimensional data, we im-

plement the boosting procedure in its componentwise version [10], in which only one regression coefficient is updated at each boosting step, namely the one which generates the best improvement in terms of minimization of the loss function. This componentwise boosting procedure yields a variable selection as a by-product: non-relevant variables, indeed, are not selected by the component-wise updating procedure and therefore their regression coefficients remain equal to 0.

The choice of the number of boosting steps (tuning parameter) highly influences the result of the procedure [30]. It controls both the sparsity of the final regression model (the higher number of steps, the more variables included in the model) and the amount of shrinkage applied to the regression coefficients (the more steps, the less shrinkage). A second tuning parameter, the amount of regularization applied to the least square estimator, has negligible influence as long as its magnitude is correctly set [3, 9].

Software: To implement the boosting technique we rely on the package *mboost* [28], which implements the model-based gradient boosting approach [27]. We use the default value 0.1 for the penalty term as commonly recommended [9] and apply a 10-fold cross-validation procedure to choose the number of boosting steps using the function *cvrisk*. To perform the favoring strategy within boosting we use the package *GAMBoost* [2], which actually implements a slightly different version of boosting [likelihood-based boosting, see 41]. In the linear regression case, however, the two boosting approaches provide exactly the same results if the penalty term is suitably chosen [14].

2.2.2 Lasso

The least absolute shrinkage and selection operator ("Lasso") [39] is a popular penalized regression technique based on an L_1 penalty term. Assuming that the vector $(y_1, \ldots, y_n)^\top$ and the vectors $(x_{1j}, \ldots, x_{nj})^\top$ (for $j = 1, \ldots, p$) have been centred and standardized, respectively, the estimates of $\beta_1, \ldots, \beta_p$ are obtained by minimizing

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda ||\beta||_1,$$

where λ is a tuning parameter ("penalty"). As in all regularization techniques, parameter estimates are shrunken towards 0 compared to least squares estimation. The L_1 penalty has the particularity that it makes some of the estimated coefficients exactly equal to 0, i.e. it leads to an intrinsic variable selection. Just as the result of the boosting procedure strongly depends on

the tuning parameter "number of boosting steps", the result of Lasso strongly depends on the tuning parameter λ, which controls the amount of penalty.

Software: Lasso is implemented in several R packages, including *penalized* [22] and *glmnet* [21]. In this paper we use the latter for the purpose of computational efficiency. The 10-fold cross-validation procedure used to choose the penalty parameter is implemented in the function *cv.glmnet*. For consistency with the other considered techniques, we use the λ-value which minimizes the mean cross-validation error.

2.2.3 Ridge regression

The ridge regression [26] is another penalized regression technique that uses an L_2 penalty term instead of the L_1 penalty used by Lasso. The estimates are obtained by minimizing

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda||\beta||_2^2,$$

where λ is again a penalty parameter. This type of regularization has the advantage that the estimator has a simple closed form. As for Lasso, parameter estimates are shrunken towards 0 compared to least squares estimation. However, a major difference between Lasso and ridge regression is that ridge regression does not perform any variable selection: some of estimates may be very close to 0 due to shrinkage, but they do not equal exactly 0. An advantage of ridge regression related to this property, however, is that it tends to cope better with highly correlated variables (while Lasso typically selects one of the correlated variables and sets the coefficients of the other to 0, leading to instability). As for the previous techniques, ridge regression requires choice of a tuning parameter, λ, which again regulates the amount of penalty.

Software: The two R packages implementing Lasso mentioned above also implement ridge regression. For the reasons already mentioned in the section on Lasso, we use *glmnet* and choose the λ-value that minimizes the mean 10-fold cross-validation error as implemented in *cv.glmnet*.

2.2.4 Elastic net

The elastic net technique [47], using both an L_1 and an L_2 penalties, can be seen as a compromise between Lasso and ridge regression. The estimates are

obtained by minimizing

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\{(1-\alpha)||\beta||_2^2+\alpha||\beta||_1\},$$

where α is an additional parameter controlling the respective weights of the L_1 and L_2 penalties. Ridge regression is the special case $\alpha = 0$, while Lasso is the special case $\alpha = 1$. While the L_1 penalty term leads to variable selection, the L_2 penalty term allows a better handling of correlated variables. If one does not want to restrict to the special cases Lasso and ridge regression, the parameter α has to be tuned along with λ.

Software: For the implementation of the elastic net we rely again on the R package *glmnet*. In this case, however, we need to choose two tuning parameters, λ and α. A two-dimensional grid 10-fold cross-validation procedure is therefore implemented (with equidistant points (distance 0.05) between 0 and 1 for α and the grid chosen by `cv.glmnet` for λ), to identify which instance of the pair (λ, α) results in the smallest mean cross-validation error.

2.2.5 SCAD

Another technique based on a penalized regression which aims at combining the strengths of the ridge regression with a variable selection procedure is the "smoothly clipped absolute deviation penalty" (SCAD) [18]. The estimates are obtained by minimizing

$$\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda I(||\beta||_1\leq\lambda)+\frac{(\alpha\lambda-||\beta||_1)_+}{\alpha-1}I(||\beta||_1>\lambda),$$

where α is a parameter > 2 and $(.)_+$ denotes the positive part. As can be seen from the formula, this technique uses a special penalty function which corresponds to a quadratic spline function with knots at λ and $\alpha\lambda$. The idea is to reduce the shrinkage applied to the parameters related to variables with large effect while maintaining the shrinkage for variables with effect close to 0. As with boosting, in theory this technique requires choice of two tuning parameters, α and λ, but one (α) is usually fixed in advance.

Software: We use the implementation of the SCAD technique available from the R package *ncvreg* [8]. As mentioned above, the tuning parameter α is fixed in advance, here to 3.7, the value suggested by the developers [18]. The value of the other tuning parameter, λ, is instead selected using the function *cv.ncvreg*, by minimizing the cross-validation error.

2.3 Evaluation criteria

Each combination of strategy and statistical method is evaluated, with varying correlation structures, in terms of prediction accuracy and ability to identify the significant variables. The former is evaluated by mean square prediction error (MSPE), the latter by a combination of sensitivity (true positive rate) and specificity (true negative rate).

3 Simulation design

The most important part of a simulation study is certainly the simulation design. As mentioned in Section 1, the focus of our study is the effect of the correlation among clinical variables, among molecular variables and between the two kinds of variables on the fitting of a regression prediction model. The models are compared in terms of prediction ability (MSPE) and correct identification of variables with (sensitivity) and without (specificity) an effect. Note that there is an extension to SIS, called ISIS, that is supposed to mitigate the issues related to the marginal approach of SIS, but showed strong convergence problems in our setting and thus was not further considered.

3.1 Data generation

As common in the simulation studies performed in a high-dimensional framework, we generate the data in blocks, which, in our context, represent gene pathways usually observed in the real data experiments. In particular, for each block h, $h = 1, \ldots, H$, we generate a "signal" from a multivariate Gaussian distribution,

$$(C_h, S_h) \sim N_{k_h}((\mu_c, \mu_m)^\top, \Sigma_h^2), \tag{1}$$

where k_h denotes the number of variables in the block, which is the sum of k_h^M, the number of molecular variables belonging to the block h (S_h), and k_h^C, the number of clinical variables related to these molecular variables (C_h). Moreover, μ_c contains the means of the clinical variables, that we set equal to 1 (i.e., here μ_c is a vector of 1 of length k_h^C), and μ_m contains the means of the molecular variables. In our simulation study, these means are all equal to 6, as the average value of the means for the gene expression proposed by [45].

Finally, Σ_h is the covariance matrix, constructed as a block matrix,

$$\Sigma_h = \left(\begin{array}{ccc|ccc} \sigma_c^2 & \dots & \rho_h^C \sigma_c^2 & \rho_h^B \sigma_c \sigma_m & \dots & \rho_h^B \sigma_c \sigma_m \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_h^C \sigma_c^2 & \dots & \sigma_c^2 & \rho_h^B \sigma_c \sigma_m & \dots & \rho_h^B \sigma_c \sigma_m \\ \hline \rho_h^B \sigma_m \sigma_c & \dots & \rho_h^B \sigma_m \sigma_c & \sigma_m^2 & \dots & \rho_h^M \sigma_m^2 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \rho_h^B \sigma_m \sigma_c & \dots & \rho_h^B \sigma_m \sigma_c & \rho_h^M \sigma_m^2 & \dots & \sigma_m^2 \end{array}\right),$$

where the first block is related to the clinical variables, the last to the molecular ones, and the other two regulate the relationship between the two kinds of variables. In detail:

- σ_c is the standard error for the **clinical** variables, in our simulations set equal to 0.5;
- σ_m is the standard error for the **molecular** variables, here equal to 0.65 [see also 45];
- ρ_h^C denotes the correlation among the clinical variables correlated with the h-th block;
- ρ_h^B denotes the correlation **between** the clinical and molecular variables in the block h;
- ρ_h^M denotes the correlation among the molecular variables of the block h.

To simplify the construction, we allow only one value for each of the correlation parameters (ρ_h^C, ρ_h^B and ρ_h^M) for each group. This means, for example, that all the molecular variables belonging to the group h have the same correlation with each other. It may happen that the Σ_h generated is non-positive definite: in this case, we apply the Higham's algorithm [25] to compute the closest positive definite matrix (which is used as covariance matrix instead of the original Σ_h).

Until now, the clinical variables differ from the molecular ones only for a slightly larger precision ($\sigma_c = 0.5$, $\sigma_m = 0.65$). In order to better differentiate the two kinds of variables and to make the data more realistic, we add some noise to the molecular part. In particular, from the signal S_h we generate a matrix of molecular data as

$$G_h = \exp\{S_h + M_h\} + E_h,$$

where M_h is a multiplicative noise [e.g., variation between the pixel affecting gene-expression measurements, see 44] and E_h an additive noise, representing the typical technical noise. Following [45], we generate these terms from Gaussian distributions, more precisely

$$M_h \sim N_{k_h}((0,\dots,0)^\top, \mathrm{diag}(\phi)),$$
$$E_h \sim N_{k_h}(\nu, \mathrm{diag}(\tau)),$$

where ϕ, ν and τ are vectors of length k_h with all terms equal to 0.1, 10 and 20, respectively. These values are selected as suggested in the R-package *Umpire* [44].

The final molecular data matrix G is created by concatenating all the G_h after a quick normalization process that consists of assigning to the smallest (smaller than 10) and the largest (> 16000) pseudo-observations a threshold value (10 and 16000, respectively) and performing a (natural, in this case) logarithm transformation.

Finally, we generate a response variable from a Gaussian distribution with mean

$$C\beta_c + G\beta_m,$$

and a standard deviation σ which depends on the setting. Here $C = (C_1 \dots C_H)$ is the matrix of clinical data, while β_c and β_m denote the vector of the true regression coefficients of the clinical and of the molecular variables, respectively.

3.2 Settings

In our study, we consider several simulation settings, that we describe in detail in this section. All the settings have in common the number of clinical (15) and molecular (10000) variables generated, and the block structure in which these variables are organized. In particular, we consider $H = 50$ blocks, with all but the last containing 10 molecular variables. The 50^{th} contain the remaining (9510) molecular variables, generated independently to each other. The first 5 blocks also contain some clinical variables; in particular: the first and the second blocks contain 3 clinical variables, the third and the fourth 2 and the fifth 5. The remaining blocks include only molecular data. Moreover, in each setting, we assign a value of ± 3 to the true regression coefficients of 5 clinical variables (relavant variables), while for the molecular part we assign a value of ± 3 to the true regression coefficients of 3 (strong effect) variables, ± 2 to those of 5 (medium effect) variables and ± 1 to the true regression coefficients of 20 (weak effect) variables. A schematic summary can be found in Table 2. The decision of keeping fixed the values

of true regression coefficient is coherent with the choice of focusing on the effect of correlation, the goal of this paper.

	Parameters	
block	β_c	β_m
$h = 1$	$(3, -3, 0)$	$(3, -3, 1, -1, 1, 0, 0, 0, 0, 0)$
$h = 2$	$(3, 0, 0)$	$(2, -2, 2, 0, 0, 0, 0, 0, 0, 0)$
$h = 3$	$(-3, 0)$	$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$h = 4$	$(0, 0)$	$(-3, 2, 1, 0, 0, 0, 0, 0, 0, 0)$
$h = 5$	$(3, 0, 0, 0, 0)$	$(-1, -1, -1, -1, -1, 0, 0, 0, 0, 0)$
$h = 6$	-	$(-2, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$h = 7$	-	$(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$
$h = 8$	-	$(1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$h = 9 - 49$	-	$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$
$h = 50$	-	$(0, \ldots, 0)$

Table 2: true effects of the clinical and molecular variables. The last block contains the regression coefficients (all equal to 0) of the remaining 9510 molecular variables.

The parameters related to the correlation, ρ_h^C, ρ_h^B and ρ_h^M, instead, assume different values within the groups in the different simulations settings, depending on the magnitude of the correlation that we want to investigate. Table 3 shows the strength of correlation in the different settings. In particular, we use the values 0.5 and 0.8 to simulate weak and strong correlation, respectively. For easy of reporting, in the following we focus on the three "homogeneous" cases, namely settings 1, 2 and 9, in which $\rho_h^C = \rho_h^B = \rho_h^M$, with $\sigma = 6$ (that corresponds to an R^2 around 0.5) and sample size $n = 500$. Results for different correlation structures, standard deviations and sample sizes are available in the Supplementary Material.

MSPE, sensitivity and specificity are computed on a large (10000 observations) independent test set generated with the same design of the training set. The results reported in Section 5 are based on 500 replications, those in the Supplementary Material on 100 replications.

4 Illustration of key issues in a simplified example

Before evaluating the effect of the correlation structure on the combinations strategies/statistical methods, it may be useful to illustrate its influence on

setting	ρ^C	ρ^M	ρ^B	situation
1	**0.0**	**0.0**	**0.0**	**no correlation at all**
2	**0.8**	**0.8**	**0.8**	**strong correlation**
3	0.5	0.8	0.8	moderate correlation among clinical
4	0.8	0.5	0.8	moderate correlation among molecular
5	0.5	0.5	0.8	strong correlation between clinical and molecular
6	0.8	0.8	0.5	moderate correlation between clinical and molecular
7	0.5	0.8	0.5	strong correlation among molecular
8	0.8	0.5	0.5	strong correlation among clinical
9	**0.5**	**0.5**	**0.5**	**moderate correlation**

Table 3: Simulation settings: different strengths of correlation in different settings. The results for the settings in bold are reported in the paper, the rest can be found in the Supplementary Material.

the estimation of the regression coefficients and the R^2. For illustration purposes, we show the results of a simplified example, in which we generate a large number of observations (5000) from a design similar to that described in Section 3.2, with the only modification concerning the total number of molecular predictors, reduced to 2000 in order to be able to implement a least squares estimator. In contrast to the main results of the paper, and due to the solely illustrative purpose of the example, we do not consider a test set, and all quantities are computed on the 5000 observations generated.

The results are presented in Table 4. For four scenarios (with and without an effect of molecular variables, with and without correlations between clinical and molecular variables) we present the estimates of the regression coefficients of some of the 15 clinical variables. In addition we report the value of the R^2 of the models using only clinical, only molecular or both kinds of variables. Here the R^2 is "adjusted" to penalize higher number of variables, using the default R implementation `adj.r.squared` of `summary.lm`,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\frac{n-1}{n-p-1} \tag{2}$$

where $\hat{y}_i$ is the estimate $\sum_{j=1}^{p}\hat{\beta}_j x_{ij}$ and $\bar{y}$ the simple average $n^{-1}\sum_{i=1}^{n} y_i$. Regarding the regression coefficient estimates, we are computing them from the "correct" model only when $\beta_{\text{mol}} = 0$, i.e., in scenarios B and D). In scenario B the parameter estimates vary around their true value, with an R^2 of 0.24. The R^2 of the molecular part should be 0, but it is slightly inflated (0.01), an issue known for several decades [33]. In the combined model (that including all 2000 irrelevant molecular variables) the R^2 is again slightly

Scenarios		$\rho = 0$				$\rho = 0.8$			
		A: $\beta_{\text{mol}} \neq 0$		**B**: $\beta_{\text{mol}} = 0$		**C**: $\beta_{\text{mol}} \neq 0$		**D**: $\beta_{\text{mol}} = 0$	
	β_{true}	est	se	est	se	est	se	est	se
β_1	3	3.13	0.22	3.26	0.17	3.55	0.46	3.27	0.31
β_2	-3	-2.97	0.22	-2.91	0.17	-2.74	0.46	-3.12	0.32
β_3	0	-0.01	0.23	0.03	0.17	0.16	0.46	-0.23	0.32
β_4	3	2.94	0.22	2.80	0.17	3.67	0.45	2.75	0.31
β_5	0	-0.02	0.22	-0.07	0.17	0.71	0.46	0.02	0.31
β_6	0	0.06	0.22	-0.11	0.17	0.90	0.46	-0.07	0.31
β_7	-3	-3.01	0.23	-2.95	0.17	-2.37	0.42	-2.74	0.29
β_8	0	-0.05	0.22	-0.13	0.17	-0.13	0.41	-0.11	0.28
β_9	0	0.03	0.22	-0.19	0.17	-3.33	0.42	-0.00	0.29
β_{10}	0	-0.32	0.23	-0.16	0.17	-3.57	0.42	0.07	0.29
β_{11}	3	3.25	0.23	3.07	0.17	2.02	0.50	3.14	0.34
β_{12}	0	0.09	0.23	0.25	0.17	-0.90	0.50	0.53	0.34
β_{13}	0	-0.04	0.22	0.03	0.17	-0.56	0.50	0.05	0.34
β_{14}	0	-0.23	0.22	-0.04	0.17	-1.67	0.49	-0.08	0.33
β_{15}	0	-0.22	0.23	-0.24	0.17	-1.95	0.50	-0.55	0.34
R^2 clin		0.16		0.24		0.23		0.17	
R^2 omic		0.37		0.01		0.61		0.11	
R^2 both		0.52		0.25		0.64		0.17	

Table 4: Estimation of the regression parameters and the R^2 in a linear regression case. The results (except for those in the last 2 rows) are based on the clinical model only. Here $\rho = \rho^C = \rho^M = \rho^B$. The heading $\beta_{\text{mol}} \neq 0$ denotes the cases in which there actually is information on the molecular part (which the clinical model cannot capture), $\beta_{\text{mol}} = 0$ denotes that there is none.

inflated (0.25 instead of the expected 0.24). Conversely, in scenario A the molecular variables have an influence on the outcome and the corresponding R^2 is severely increased when correctly including both parts in the model (0.52). But for a small random difference, it is the sum of the R^2's related to the only clinical ($R^2 = 0.16$) and only molecular ($R^2 = 0.37$) models. Note that the standard errors of the estimates increase from scenario B to scenario A. Excluding relevant molecular variables from the model also leads to a larger estimation of the standard errors for the regression coefficients of the clinical variables, actually independent of whether the former and latter variables are correlated (the same concept works when comparing scenarions C and D). Since

$$\mathrm{Var}(\hat{\beta}) = (X^T X)^{-1}\sigma^2,$$

its estimation (and that of its square root) depends on the estimation of the residual variance,

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Since the response actually depends on variables not included, the model cannot provide correct estimates, i.e., the values $\hat{y}_i$ are farther with respect to y_i than the cases in which y_i does only depend on the variables included in the model. Therefore we have a larger estimation of the residual variance.

This also explains why the R^2 for the clinical model is smaller when there is actually information on the molecular variables, even if clinical and molecular variables are uncorrelated (contrast the R^2 clin's in scenarios A and B). It is again an effect of the larger prediction error, as it can be seen in formula (2), where the quantity $(y_i - \hat{y}_i)^2$ is now in the numerator.

In scenarios C and D we added substantial correlation (both among clincial/molecular variables and between the two kinds of variables). If the molecular variables have no effect (scenario D) the model for the clinical part has more uncertainty (larger standard errors and smaller R^2 compared to scenario B), but parameter estimates are still unbiased. In scenario C, in contrast, the molecular variables have an effect and estimates of the "incorrect" clinical model are (substantially) biased. This is most obvious for the estimates of β_9 and β_{10} (correlated with the molecular variable $31, 32$ and 33, see Table 2) and, to a smaller extent, $\beta_{12} - \beta_{15}$ (correlated with the molecular variable $41 - 45$). The parameter estimate bias for scenario C is certainty the most evident characteristic of the table. When relevant variables (in this example, the molecular variables with an effect) are not considered in the model, their effect is caught by the regression coefficients of those included variables which are correlated with them, causing bias. As expected, in the

other three scenarios there is no bias, either because the model is "correct" (D), there is no correlation (B) or both (A).

Finally the R^2 for the clinical model is the highest in scenario C, again because the clinical variables, due to strong correlation, also capture information belonging to the molecular ones. Conversely, the R^2 in scenario D is smaller than that in scenario B (although in both cases there is no information in the molecular variables and so we are estimating from the "correct" model) because the clinical variables can explain less variability due to the high correlation (basically all of them contain the same – very similar – information).

5 Results

As already mentioned in the results part of the simulation profile we start by concentrating on MSPE. Note that for better visualization we decided to exclude all combinations with SIS and ridge from the plots, because the related prediction error are not comparable to the other. A detailed explanation is given in the Discussion. The numerical results are nevertheless reported in Tables 5, 6 and 7 (without SIS).

5.1 Prediction error

Tables 5, 6 and 7 report the results in terms of MSPE in the three scenarios under investigation (no correlation, high correlation and moderate correlation, respectively). The column and the row corresponding to the best performance (i.e., smallest MSPE) are highlighted in bold. The results for these columns and rows are also reported in Figure 1.

Correlation. In general, we can affirm that the strength of the correlation seems to highly affect the results. Larger correlation means a smaller prediction error (see Figure 2). It is most probably a consequence of the fact that, with increasing correlation, it is less critical if a variable with influence is falsely excluded from the final model, as its effect is "taken over" by correlated variables included in the model. In fact, this pattern (more correlation, smaller MSPE) is true but, not surprisingly, for the ridge regression. As in ridge regression all variables are included in the model, there is no advantage in having correlated covariate, and the classical phenomenon of better estimates in case of no correlation holds.

[Figure 2 approximately here]

	strategy						
statistical	naive	clinical offset		**favoring**		dim. reduct.	
method	1	2.1	2.2	3.1	**3.2**	4.1	4.2
scad	**54.76** (2.97)	**54.13** (2.65)	**53.46** (2.63)	**53.38** (2.36)	**52.93** (2.32)	**59.30** (3.57)	**60.84** (3.76)
lasso	56.35 (2.24)	56.08 (2.05)	55.37 (2.09)	55.14 (2.03)	**54.67** (1.99)	57.04 (4.27)	60.42 (4.41)
elastic net	56.72 (2.39)	56.31 (2.20)	55.65 (2.22)	55.39 (2.20)	**54.94** (2.06)	58.48 (5.03)	62.09 (5.26)
ridge	91.50 (0.86)	66.21 (0.85)	65.68 (0.87)	65.92 (0.78)	**65.47** (0.79)	96.69 (0.88)	96.68 (0.84)
boosting	57.26 (2.20)	56.07 (1.92)	55.39 (1.96)	55.44 (2.00)	**54.93** (1.95)	54.98 (2.41)	58.06 (2.11)

Table 5: setting 1 (no correlation), MSPE for all combinations strategies/statistical methods.

Strategies and correlation. There does not seem to be many differences among the strategies. However, naive (1) and dimensionality reduction with the reduced model (4.2) almost always have the two worst performances. Dimensionality reduction with the full model (4.1) is the only strategy which seems strongly affected by the amount of correlation, as it is among the worst in case of no correlation, among the best in case of correlation. It is worth noting, however, that in the latter case its performance is very close to that of favoring (both with reduced and full clinical model) for the best statistical methods (lasso, elastic net and boosting), while worse in combination with SCAD and, very severely, with ridge regression. In contrast, the favoring strategy (both considering the full and the reduced clinical models) is always close to the best, if not the best, result, no matter the amount of correlation.

Statistical methods and correlation. Lasso, boosting and elastic net have, not really surprisingly, very similar results, no matter the amount of correlation. In addition, their performances are always among the best. SCAD, instead, seems to perform really well in the case of uncorrelated variables, and slightly worse than the three aforementioned methods in case of correlation. Finally, ridge regression has always the worst results. Careful considerations of the simulation settings, reported in the discussion, can explain this poor performance.

[Figure 1 approximately here]

	strategy						
statistical	naive	clinical offset		favoring		**dim. reduct.**	
method	1	2.1	2.2	3.1	3.2	**4.1**	4.2
scad	52.62 (2.48)	53.58 (2.39)	52.86 (2.43)	51.53 (2.16)	51.42 (2.22)	**54.56** (2.25)	56.39 (2.45)
lasso	48.41 (1.78)	49.07 (1.91)	48.34 (2.07)	46.66 (1.55)	46.70 (1.71)	**45.83** (1.93)	48.98 (1.96)
elastic net	48.58 (1.84)	49.23 (2.10)	48.47 (2.20)	46.83 (1.69)	46.81 (1.85)	**46.17** (2.39)	49.39 (2.36)
ridge	96.10 (1.69)	78.52 (1.55)	77.84 (1.69)	77.18 (1.41)	76.85 (1.53)	**94.81** (1.87)	94.88 (1.85)
boosting	**48.81** (1.68)	**48.83** (1.85)	**48.09** (2.00)	**46.83** (1.51)	**46.75** (1.68)	**44.92** (1.15)	**48.09** (1.33)

Table 6: setting 2 (high correlation), MSPE for all combinations strategies/statistical methods.

5.2 Sensitivity and specificity

In addition to the MSPE, we evaluate the ability of the combination strategies/statistical methods to identify the relevant variables. We report the results corresponding to the rows highlighted in bold in Tables 5, 6 and 7 (i.e., for the best statistical method) in Figures 3 and 4. The former contains information on sensitivity and specificity for the clinical variables, the latter for the molecular variables. In Figure 5 the same quantities, related to the molecular variables, are reported for the favoring (with reduced model) strategy, which is the best in the uncorrelated setting and has comparable (sometimes even better) results in the correlated settings (it does not make sense to consider the dimensionality reduction strategy, as the molecular variables are summarized into a single score).

[Figure 3 approximately here]

[Figure 4 approximately here]

[Figure 5 approximately here]

5.2.1 Clinical data

We have 15 clinical variables, five of which have an effect on the outcome. For the three versions without variable selection (strategies 2.1, 3.1 and 4.1) results are by definition 1 (sensitivity) or 0 (specificity) and will be therefore ignored in the following. Note that the variable selection procedure is common for strategies 2.2, 3.2 and 4.2 (backward elimination), so the contrast

	strategy						
statistical	naive	clinical offset		favoring		**dim. reduct.**	
method	1	2.1	2.2	3.1	3.2	**4.1**	4.2
scad	55.96 (2.56)	54.55 (2.33)	53.80 (2.43)	53.03 (2.22)	52.65 (2.34)	**56.70** (3.43)	59.17 (3.29)
lasso	53.36 (2.25)	53.25 (2.24)	52.47 (2.35)	51.36 (2.07)	51.10 (2.14)	**50.70** (3.02)	54.70 (2.92)
elastic net	53.63 (2.33)	53.50 (2.43)	52.65 (2.47)	51.53 (2.10)	51.33 (2.26)	**51.42** (3.43)	55.47 (3.24)
ridge	97.95 (1.33)	77.02 (1.28)	76.38 (1.40)	76.44 (1.15)	76.01 (1.28)	**99.44** (1.46)	99.47 (1.43)
boosting	**54.78** (2.07)	**53.25** (2.15)	**52.45** (2.25)	**51.53** (1.99)	**51.17** (2.08)	**49.09** (1.74)	**53.16** (1.88)

Table 7: setting 9 (moderate correlation), MSPE for all combinations strategies/statistical methods.

is fundamentally between them and the naive strategy. Figure 3 supports the statement "clinical variables risk to get lost among the molecular variables" [3, 4] if not adequately treated. Especially in case of correlation (first column, second and third rows) the percentage of relevant clinical variables correctly identified is relatively small, in median less than 50% in the case of strong correlation (setting 2). On the other hand, hardly any irrelevant clinical variable is included in the model (right column). Actually, for the situation with no correlation (setting 1) sensitivity and specificity for the naive approach are close to 1, but this excellent result is not relevant for practical applications. For the two settings with correlation (settings 2 and 9), the backward elimination procedure applied to the clinical data leads to median sensitivity and specificity around 0.8, indicating that in most cases variables were correctly included or excluded. For the naive strategy, instead, it is between 0.4 (strong correlation) and 0.6 (moderate correlation).

5.2.2 Molecular data

In Figures 4 and 5 we show sensitivity and specificity of the models only concerning the molecular data. Sensitivity relates to the inclusion of the 28 variables with an effect on the outcome, whereas specificity gives the rate of correct exclusions of the 9972 variables without an effect. Consequently, a specificity of 99% means that about 100 variables without effect were included in the model. In Figure 4 we compare results for the different strategies. Concerning sensitivity, differences are negligible (note that the two reduction strategies 4.1 and 4.2 reduce the molecular information to a unidimensional score and are therefore not relevant here). Note that the differences among

strategies are minimal, and a small advantage for the naive strategy in setting 2 is the only relevant difference. Altogether, results clearly demonstrate that a substantial part of the 28 relevant variables are not identified by any of the strategies. More severe differences exist for the exclusion of variables without any effect. For setting 1 (no correlation) the results are very similar for all strategies (all with SCAD), with all distributions centred around 0.99, implying that about 100 irrelevant variables were included. Results are a bit worse for the naive strategy but the difference is not that large. For the two settings with correlated variables, however, the specificity of naive seems slightly better (around 0.998) than for the other. The median of the favoring approach is a bit lower and in some cases specificity is below 0.99, which means that more than 100 variables without effect were incorrectly included. Corresponding plots comparing selection strategies (Figure 5) indicate more differences. Concerning sensitivity, SCAD is slightly better than others if variables are uncorrelated, but it has much lower values in the settings with correlation. Sensitivity of lasso, elastic net and boosting are similar. For the two correlated settings, elastic net has slightly lower specificity but altogether differences are not that large (the values for ridge regression are not reported in the plot as they are, by definition, 0).

5.3 Pre-screening

Some authors [e.g., 19] advise applying a pre-screening step which reduces the number of irrelevant variables before actually implementing the main statistical method. In this simulation study we implemented two well-known procedures, the "sure independent screening" (SIS) and the "iterative sure independent screening" (ISIS), both by [19]. The former is basically a univariate method which is supposed to remove those variables which have no association with the outcome. The marginal correlation between each variable and the outcome is computed, and the d variables with the highest correlation are kept. The latter procedure is a modification of the former which aims at handling possible spurious correlation/multicollinearity issues. In ISIS the variables are iteratively allowed to enter and exit the list of the relevant ones, based on their correlation with the excluded (by the previous step) variables and with the residuals of a model fitted using the variables selected in the previous step.

In this simulation study. Both methods did not work very well in our simulation study. ISIS, in particular, never converged, even when allowing many more (ten times) iterations than the default value (10). Also considering the huge amount of computational time necessary to run it (probably

not really problematic in a single study, an issue for extensive simulations), we decided to exclude this procedure from the study. Results for SIS (here implemented fixing $d = 1000$), instead, have not been reported in the analyses above due to their bad performances. Figure 6 reports the results in one specific case (in the case of strong correlation – setting 9 – but the results are similar in the other scenarios). The only case in which SIS improves the prediction is when ridge regression is applied. This can be easily explained by the necessity of reducing the variance by removing the variables without any effect from the prediction model. Ridge regression is, indeed, the only statistical method here considered without intrinsic variable selection.

[Figure 6 approximately here]

6 Illustration on a real dataset

Here, we consider the data from [13] (available at the EMBL-EBI ArrayExpress database, `www.ebi.ac.uk/arrayexpress`, under accession number E-GEOD-33070). They contain information about the percent of body weight change of 26 kidney transplant recipients, together with some clinical information (presence of diabetes, ethnicity and gender) and 28869 gene expressions. Looking at the first 10 gene expressions measured (see Table S.1 in the Supplementary Material), we note a situation similar to our second setting. Some gene expressions (e.g., the 1st) are basically uncorrelated to the others, while other gene expressions (e.g., the 2nd, the 3rd and the 5th) form clusters of highly correlated gene expressions (the correlation between the 2nd, the 3rd and the 5th is around 0.8). It can be noted, however, that the correlation structure is more complicated than in our scenarios: we discuss the simulation design simplifications in Section 7.

In contrast to our settings, here the clinical variables are dichotomous. Some are obviously correlated (e.g., diabetes before and after an operation), some clearly not (see *diabetes-pre-operation* and *ethnicity*). About the relation with the gene expressions, let us consider as an example *gender*: Loosely using the point-biserial correlation coefficient [17, it is impossible to check the assumptions with only 26 observations], we can see that it is independent of some gene expressions (e.g., 1st and 4th gene expression) and strongly dependent on others (e.g., 8th).

Applying a specific split in training and test set, using boosting as a statistical method we obtained a MSPE for the naive strategy equal to 58.31, larger than the 32.69 obtained with strategy 4.1. Note that the sample size is very small, so every split can lead to very different results (adding 10 more

replications, we obtained an average MSPE of 62.50 and 61.51, respectively, suggesting that there are splits in which strategy 4.1 performed even worse than the naive strategy). Even with a larger sample size, however, one cannot draw conclusions from a single (or a few) dataset(s) (it would mean a study with $n = 1$), so this section must be understood only as an illustration.

Finally, since we do not know the truth, nothing can be said about sensitivity and specificity.

7 Discussion

The results of this simulation provide some insight into the problem of interest, confirming some well-known concepts (the necessity of treating the clinical variables in a suitable way, the impact of the strength of the correlation on the results, ...) and providing some unexpected evidence (e.g., summarizing all molecular information in a score to be added to the clinical model can provide very good results). The limitations of this study, however, are quite clear, as only two values for the correlation have been investigated, only strong effects on the clinical part have been considered, and only default procedures to find the tuning parameters for the statistical methods have been implemented.

In any case, one should be aware that a simulation study is limited by nature, and, no matter how much "realistic" the data generating process is, reality can be very different. Simplifications and specific choices for several details are unavoidable, leading to a certain amount of arbitrariness [see e.g. 6]. An important exercise is to analyse the impact of the simulation design's characteristics on the results, in order to get useful information.

Sparse scenario. In this simulation study we only considered a limited number of relevant variables (5 clinical and 25 molecular) and a large quantity of completely irrelevant variables. This is most probably the reason why ridge regression performs so poorly in our study, especially in contrast to lasso. Tibshirani [39] showed that in the low-dimensional setting lasso has better performances than ridge regression when there are only a few variables with strong effect which carry the information. The contrary is true when there are many relevant variables with small effects. Since ridge regression has relatively good performance in many real data studies [see, e.g., 7], where the signal may come from several low-effect variables, it is most likely that the behaviour described in [39] also applies to high-dimensional settings.

Probably for the same reason, elastic-net seems to perform constantly worse than lasso, although sometimes not by much. If the data are simu-

lated in a way that is favorable to lasso, elastic-net can only get close to its performance by heavily weighting the L_1 term in its mixture penalty. Figure 4, which shows very similar sensitivity and specificity for lasso and elastic-net, seems to support this statement.

Linear and additive effects. Here we considered only linear and additive effects. While this is probably not realistic, the assumption is often made in real data studies. It is worth noting that this characteristic may be highly penalizing for pre-selection techniques like SIS and ISIS. A pre-selection procedure like SIS may be highly effective, for example, in the case of multi-modal marginal distributions in the case of classification of clustering [see, e.g., 15], while it seems unable to discriminate between relevant and irrelevant variables in our simulation study. Again, considerations driven from this simulation study must be taken as design-specific and not true in general.

Simplified correlation structure. Another strong deviance from a realistic scenario is the regularity of our correlation structure. On the one hand, this simplification (we have 50 blocks and for all blocks we have one single value for the correlation among clinical variables, one for the correlation among the molecular variables and one for the correlation between the two) really helps understanding of the influence of the correlation strength on the results. On the other hand, one would never experience a situation like this in a real dataset. We saw in Section 6 that the correlation structure, even limited to the first 10 gene expressions, is more complex. If, in contrast to us, one is not interested in controlling the correlation structure, one may prefer to use real data and only simulate the outcome [11].

Note that, with regard to our study, a simplified correlation structure may represent an additional advantage for lasso in comparison to, for example, ridge regression, as the latter is better than the former in dealing with correlation among explanatory variables. Componentwise boosting also "bets" on data sparsity.

Use of software default parameters. None of the statistical methods has been optimized for the specific simulation, and our computations largely rely on the default procedures implemented in the R packages. On the one hand, this choice guarantees a fair comparison [43]. Familiarity with a specific method, indeed, may result in its better optimization for the problem on hand and, consequently, a competitive advantage against other approaches [see, e.g., 5]. On the other hand, the use of default values does not allow

fully exploitation of the potentialities of a statistical method. For example, in our simulations, the value of the number of boosting steps, the most important tuning parameter for boosting, was often selected as the largest value (100) allowed by the default settings of the R routine used to perform cross-validation. This issue may have had an impact on the results, for example by favoring the dimensionality reduction strategy. In fact, when combined with boosting, this approach performed better than the favoring strategy, in contrast, for example, to what happens when using the usually similarly-performing lasso (see, e.g., Table 6). Not having to update the regression coefficients of the clinical variables, boosting could allocate all the iterations to update the molecular variables and provide a very good molecular score.

Another method which could have been penalized from the limited space in which the tuning parameters have been investigated is elastic-net. For the mixing parameter α we used a grid with intervals of 0.05. If the best value would have been, let us say, 0.03, our tuning procedure would have selected 0.05. The results of elastic-net might have been closer to those of lasso if a finer grid had been used.

The choice of the optimal tuning parameter in the context of data integration is an issue that certainly deserves further investigation. It is known to be a key aspect of the method implementation [see, e.g. 20, 29, 30, 38], which should take into consideration the goal of the analysis [24]. On the other hand, excessive tuning may mean a reduced generalizability of the results, as the choice of the tuning parameters may be driven by specific characteristics of the dataset on hand [34].

Finally, note that for some tuning parameters the use of default values is recommended by the authors of the methods themselves. This is the case for the parameters $\alpha(= 3.7)$ of SCAD and the boosting step size $\nu(= 0.1)$ of boosting. As long as their magnitude is reasonable, these parameters do not affect the performance of the methods too much [18, 9]. There is no point, therefore, to change them.

Evaluation criteria. Finally, we used classical criteria to evaluate the performances of the combination strategies/statistical methods. A different combination strategy/statistical method might have resulted the best when using an absolute distance instead of a quadratic distance in the computation of the prediction error. More importantly, sensitivity and specificity, although standard choices in biomedical applications, do not fully capture the capacity of a combination strategy/statistical method to select the best variables for prediction purposes. In particular, the selection of a variable

correlated to a relevant one should be accredited when the former can easily substitute the latter in the model. Criteria to capture this aspect, for example that of [23], are worth investigation. Further discussion on this topic can be found, among others, in [31],[36] and [37].

Final remarks. In this study we started by comparing 105 combinations of strategies, statistical methods and pre-screening steps, to derive a combined model. Since in our simulations ISIS did not converge, the description in this paper only involves 70 of them. Furthermore we consider several settings concerning the correlation structure. We consider this as a first study to provide a general overview of some strengths and weaknesses of the considered approaches. Despite the limitations of the simulation design, significant insights into the suitability of combination strategies/statistical methods and early ideas about their advantages and disadvantages can be driven. Further studies may start from these findings to perform a more detailed comparison of the most promising combinations identified here.

Key Points

- Combining low-dimensional clinical and high-dimensional molecular information in a prediction model is beneficial but there are difficulties, including handling complex correlation structures.
- Seven strategies to combine clinical and molecular variables and five methods to derive a prediction model are contrasted, with and without a pre-selection step, for a total of seventy strategy/method/screening combinations.
- Depending on the correlation structure, specific combinations provide better results in terms of prediction ability and selection of the relevant variables.

Acknowledgement

We thank Tim Haeussler for technical help, Alethea Charlton and Kaya Miah for language corrections. DBR was initially supported by grant BO3139/4-2 to ALB and WS was supported by grant SA580/8-2. Both grants are from the German Research Foundation (DFG). Computations were performed on the

Abel Cluster, owned by the University of Oslo and Uninett/Sigma2 (project NN9480K).

Biographical Note

Riccardo De Bin is an associate professor at the Department of Mathematics, University of Oslo (Norway). His research is at the interface between computational, theoretical and applied statistics.

Anne-Laure Boulesteix is an associate professor at the Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich (Germany). Her research mainly focuses on computational statistics, biostatistics and research on research methodology.

Axel Benner is a scientist at the Division of Biostatistics of the German Cancer Research Centre of Heidelberg (Germany). His research interests include time-to-event data analysis, statistics for high-dimensional covariate space and clinical statistics.

Natalia Becker is a post-doc at the Division of Biostatistics of the German Cancer Research Centre of Heidelberg (Germany). Her research interests include chemometrics, clinical statistics and survival analysis.

Willi Sauerbrei is a professor at the Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg (Germany). His research interests span both methodological statistics and clinical applications.

References

[1] D. G. Altman, L. M. McShane, W. Sauerbrei, and S. E. Taube. Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Medicine*, 10:51, 2012.

[2] H. Binder. *GAMBoost: Generalized linear and additive models by likelihood based boosting*, 2013. URL `http://CRAN.R-project.org/package=GAMBoost`. R package version 1.2-3.

[3] H. Binder and M. Schumacher. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9:14, 2008.

[4] A.-L. Boulesteix and W. Sauerbrei. Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics*, 12:215–229, 2011.

[5] A.-L. Boulesteix, R. Wilson, and A. Hapfelmeier. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*, 17:138, 2017.

[6] A.-L. Boulesteix, H. Binder, M. Abrahamowicz, W. Sauerbrei, and Simulation Panel of the STRATOS Initiative. On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, 60:216–218, 2018.

[7] H. Bøvelstad, S. Nygård, and Ø. Borgan. Survival prediction from clinico-genomic models - a comparative study. *BMC Bioinformatics*, 10:413, 2009.

[8] P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5:232–253, 2011.

[9] P. Bühlmann and T. Hothorn. Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22:477–505, 2007.

[10] P. Bühlmann and B. Yu. Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.

[11] A. Burton, D. G. Altman, P. Royston, and R. L. Holder. The design of simulation studies in medical statistics. *Statistics in Medicine*, 25: 4279–4292, 2006.

[12] V. Canuel, B. Rance, P. Avillach, P. Degoulet, and A. Burgun. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. *Briefings in Bioinformatics*, 16:280–290, 2014.

[13] A. Cashion, A. Stanfill, F. Thomas, L. Xu, T. Sutter, J. Eason, M. Ensell, and R. Homayouni. Expression levels of obesity-related genes are associated with weight change in kidney transplant recipients. *PloS ONE*, 8:e59962, 2013.

[14] R. De Bin. Boosting in Cox regression: a comparison between the likelihood-based and the model-based approaches with focus on the R-packages CoxBoost and mboost. *Computational Statistics*, 31:513–531, 2016.

[15] R. De Bin and D. Risso. A novel approach to the clustering of microarray data via nonparametric density estimation. *BMC Bioinformatics*, 12:49, 2011.

[16] R. De Bin, W. Sauerbrei, and A. L. Boulesteix. Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in Medicine*, 33:5310–5329, 2014.

[17] F. Drasgow. Polychoric and polyserial correlations. In *The Encyclopedia of Statistics*, volume 7, pages 68–74. Wiley, 1986.

[18] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360, 2001.

[19] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Methodological)*, 70:849–911, 2008.

[20] Y. Fan and C. Y. Tang. Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:531–552, 2013.

[21] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1, 2010.

[22] J. Goeman, R. Meijer, and N. Chaturvedi. *penalized: L1 (lasso and fused lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*, 2014. URL `http://CRAN.R-project.org/package=penalized`.

[23] M. G. G'Sell, T. Hastie, and R. Tibshirani. False variable selection rates in regression. *arXiv preprint arXiv:1302.2303*, 2013.

[24] K. H. Hellton and N. L. Hjort. Fridge: Focused fine-tuning of ridge regression for personalized predictions. *Statistics in Medicine*, 37:1290–1303, 2018.

[25] N. J. Higham. Computing a nearest symmetric positive semidefinite matrix. *Linear Algebra and its Applications*, 103:103–118, 1988.

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

[27] T. Hothorn and P. Bühlmann. Model-based boosting in high dimensions. *Bioinformatics*, 22:2828–2829, 2006.

[28] T. Hothorn, P. Buehlmann, T. Kneib, M. Schmid, and B. Hofner. *mboost: Model-Based Boosting*, 2014. URL `http://CRAN.R-project.org/package=mboost`. R package version 2.4-0.

[29] Q. Hu and C. S. Greene. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell rna transcriptomics. In *PSB*, pages 362–373, 2019.

[30] A. Mayr, B. Hofner, and M. Schmid. The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. *Methods of Information in Medicine*, 51:178–186, 2012.

[31] S. Michiels, S. Koscielny, and C. Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, 365: 488–492, 2005.

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

[33] A. C. Rencher and F. C. Pun. Inflation of r^2 in best subset regression. *Technometrics*, 22:49–53, 1980.

[34] W. Saelens, R. Cannoodt, and Y. Saeys. A comprehensive evaluation of module detection methods for gene expression data. *Nature Communications*, 9:1090, 2018.

[35] H. Seibold, C. Bernau, A.-L. Boulesteix, and R. De Bin. On the choice and influence of the number of boosting steps for high-dimensional linear cox-models. *Computational Statistics*, 33:1195–1215, 2018.

[36] R. M. Simon, J. Subramanian, M.-C. Li, and S. Menezes. Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Briefings in Bioinformatics*, 12:203–214, 2011.

[37] Y. Takwoingi, B. Guo, R. D. Riley, and J. J. Deeks. Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data. *Statistical Methods in Medical Research*, 26:1896–1911, 2017.

[38] J. Thomas, A. Mayr, B. Bischl, M. Schmid, A. Smith, and B. Hofner. Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 28:673–687, 2018.

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

[40] C. Truntzer, E. Mostacci, A. Jeannin, J.-M. Petit, P. Ducoroy, and H. Cardot. Comparison of classification methods that combine clinical data and high-dimensional mass spectrometry data. *BMC Bioinformatics*, 15:385, 2014.

[41] G. Tutz and H. Binder. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62:961–971, 2006.

[42] C. D. van Karnebeek, S. B. Wortmann, M. Tarailo-Graovac, M. Langeveld, C. R. Ferreira, J. M. van de Kamp, C. E. Hollak, W. W. Wasserman, H. R. Waterham, R. A. Wevers, et al. The role of the clinician in the multi-omics era: are you ready? *Journal of Inherited Metabolic Disease*, 41:571–582, 2018.

[43] L. M. Weber, W. Saelens, R. Cannoodt, C. Soneson, A. Hapfelmeier, P. P. Gardner, A.-L. Boulesteix, Y. Saeys, and M. D. Robinson. Essential guidelines for computational method benchmarking. *Genome Biology*, 20:125, 2019.

[44] J. Zhang and K. Coombes. UMPIRE: Ultimate Microarray Prediction, Inference, and Reality Engine. In *BIOTECHNO 2011, The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies*, pages 121–125, 2011.

[45] J. Zhang, P. Roebuck, and K. Coombes. Simulating gene expression data to estimate sample size for class and biomarker discovery. *International Journal on Advances in Life Sciences*, 4:44–51, 2012.

[46] B. Zhu, N. Song, R. Shen, A. Arora, M. J. Machiela, L. Song, M. T. Landi, D. Ghosh, N. Chatterjee, V. Baladandayuthapani, et al. Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Scientific Reports*, 7:16954, 2017.

[47] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, 67:301–320, 2005.

Figure 1: MSPE for the best statistical method (left column) and best strategy (right column) in setting 1 (top row), 2 (middle row) and 9 (bottom row). Results for ridge regression have been excluded for a better visualization.

Figure 2: Summary of the MSPE obtained with the several combinations strategy/statistical method (excluding ridge) for the three selected scenarios (coefficient of correlation between brackets). The box-plots report the average MSPE for all combinations (ridge excluded) in 500 replications.

Figure 3: clinical data, sensitivity (left column) and specificity (right column) for SCAD in setting 1 (first row), boosting in setting 2 (second row) and boosting in setting 9 (third row).

Figure 4: molecular data, sensitivity (left column) and specificity (right column) for SCAD in setting 1 (first row), boosting in setting 2 (second row) and boosting in setting 9 (third row).

Figure 5: molecular data, sensitivity (left column) and specificity (right column) for the favoring (with reduced clinical model) strategy in setting 1 (first row), setting 2 (second row) and setting 9 (third row). Sensitivity and specificity for ridge regression are outside the plot limits, always equal to 1 and 0, respectively.

Figure 6: MSPE for the different combinations strategy/statistical method with (grey) and without (white) a pre-selection step based on SIS.