LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Riccardo De Bin, Willi Sauerbrei, Anne-Laure Boulesteix

Investigating the prediction ability of survival models based on both clinical and omics data: two case studies

Technical Report Number 153, 2014
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Investigating the prediction ability of survival models based on both clinical and omics data: two case studies

Riccardo De Bin* Willi Sauerbrei †
Anne-Laure Boulesteix ‡

February 18, 2014

Abstract

In biomedical literature numerous prediction models for clinical outcomes have been developed based either on clinical data or, more recently, on high-throughput molecular data (omics data). Prediction models based on both types of data, however, are less common, although some recent studies suggest that a suitable combination of clinical and molecular information may lead to models with better predictive abilities. This is probably due to the fact that it is not straightforward to combine data with different characteristics and dimensions (poorly characterized high dimensional omics data, well-investigated low dimensional clinical data). In this paper we analyze two publicly available datasets related to breast cancer and neuroblastoma, respectively, in order to show some possible ways to combine clinical and omics data into a prediction model of time-to-event outcome. Different strategies and statistical methods are exploited. The results are compared and discussed according to different criteria, including the discriminative ability of the models, computed on a validation dataset.

*debin@ibe.med.uni-muenchen.de - Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität of Munich, Germany

†wfs@imbi.uni-freiburg.de - Department of Medical Biometry and Medical Informatics (IMBI), University Medical Center Freiburg, Germany

‡boulesteix@ibe.med.uni-muenchen.de - Department of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität of Munich, Germany

1 Introduction

In the last 15 years, the progress in the generation of high-throughput molecular data (omics data) has raised high expectations in biomedical research. In particular, large scale gene expression data were generated and analysed in numerous studies, often with an emphasis on their potential to identify so-called gene signatures with the aim to predict a specific outcome of the considered disease. In recent years, however, the initial enthusiasm has been tempered, with the publication of a number of critical studies claiming the inefficacy of omics data for predictive purposes [1, 2, 3, 4, 5]. While the prediction ability of models based on omics data only is under debate, the advantage of integrating clinical data and omics data seems to be gaining consensus [6, 7, 8, 9] and is supported by recent comparative studies, in which the combined models often outperform those models based only on clinical or only on omics data [10, 11].

The different dimensions and characteristics of clinical and omics data, however, make their combination not straightforward from a statistical point of view. For example, if no particular attention is paid to the clinical variables, they can easily "get lost" in the high number of omics variables [11]. Some strategies to combine these two kinds of data have been reviewed in a recent paper by Boulesteix and Sauerbrei [12]. We follow their theoretical framework, showing how some statistical selection methods can be considered and adapted in practice in order to take simultaneously advantage of the clinical and the omics information. The methods are demonstrated through application to two publicly available datasets from to a breast cancer study [13] and a neuroblastoma study [14], respectively. The resulting models are compared in terms of goodness of prediction and discriminative ability.

The paper is structured as follow: the data are introduced in Section 2 while the statistical tools are described in Section 3: in particular, after a short overview on the predictive methods, we show how to adapt them in order to exploit the strategies reviewed in the aforementioned paper by Boulesteix and Sauerbrei [12]. The numerical results of the study, together with some comments, are presented in Section 4, while additional remarks are presented in Section 5.

2 Data

2.1 Breast cancer data

The first considered dataset was collected from patients with newly diagnosed ERBB2-negative breast cancer by Hatzis and colleagues [13] for their study on a genomic predictor of response and survival following Texane-Anthracycline chemotherapy. The censored response is the distant relapse free survival time, i.e., the time interval between the initial diagnosis biopsy and either the diagnosis of distant metastasis or the death [15]. The data can be retrieved from the publicly available Gene Expression Omnibus repository (GSE25066). This dataset has the advantage that the clinical data are also publicly available in addition to the omics information, which is (unfortunately) not common in publicly available repositories. The data are contained in two clearly separated datasets, which are used in our analyses as training and validation sets. In both sets, the number of observations is relatively large, especially if compared to similar studies. In particular, they contained information about 310 and 198 patients, respectively. However, the effective sample size, i.e. the number of observed events (non-censored observations) is much smaller (66 events in the training set and 45 in the validation set). Nevertheless, this is common in survival analysis of omics data. These numbers are larger than those of many comparable studies. Due to the missing values, some observations are excluded from our analysis, leading to training and validation sets of 283 (58 events) and 182 (41 events) observations, respectively.

In order to construct a clinical model, we selected the variables presented in Table 1 of the original paper [13], namely *age*, *nodal status*, *tumor size*, *grade*, *estrogen receptor* and *progesterone receptor*. We did not consider *AJCC stage* because it is a classification system based on *nodal status*, *tumor size* and *metastasis*. The latter is identical in all patients (no metastasis) and the two other are already considered as single predictors in our analysis. We dichotomized the variable *age* using the threshold 40 years, which seems to be more relevant for prediction than the cutpoint 50 used in the original analysis [16]. The most relevant difference between the original study and ours, however, lies in the use of all available information for *nodal status*. While in the original paper it was considered as a dichotomous variable (negative versus positive lymph nodes), we use three indicator variables to differentiate between N0 (no lymph node involved), N1 (from 1 up to 3), N2 (4-9) and N3 (10 and more). Therefore, we considered *nodal status* as an ordinal categorical variable with 4 modalities. Also *tumor size* and *grade* are ordinal categorical variables, with 4 and 3 modalities, respectively. Three of the 508

patients had a T0 tumor which we collapsed with the T1 group. Finally, both *estrogen receptor* and *progesterone receptor* are dichotomous variables which indicate whether the cancer has (or has not) the receptors for estrogen or progesterone, respectively.

2.2 Neuroblastoma data

The second considered dataset refers to the study conducted by Oberthuer and colleagues [14] on patients with neuroblastoma, a malignant pediatric tumor. The original data, available at the ArrayExpress database (accession number E-MTAB-16) contain microarray information about 9978 genes of 376 patients. The recorded outcome is the overall survival time. In our analysis, we refer only to those 362 observations included in the analysis performed by Bøvelstad and colleagues in their 2009 paper [10]. For these patients, indeed, an important clinical predictor is available: the predictor *risk group* according to the German neuroblastoma trial, which the authors consider as a dichotomous variable with levels 0 (low/intermediate risk) and 1 (high risk). The other 14 observations were excluded due to lack of clinical information. Unfortunately, by considering their version of the data, we do not have any information on the original split between training and validation set. We recover it arbitrarily by randomly splitting the observations into a training and a validation set, the former with 240 patients, the latter with 122 (the cardinality of the two sets are derived by Bøvelstad and colleagues [10]). In our split, the training set contains 51 events, the validation set 24. In addition, based on further information present in the aforementioned repository, we could also include in our study also the variable *age*, through a dichotomous variable indicating whether the patient is more or less than 444.5 days old (i.e., the median) at the time of the diagnosis.

3 Combining clinical and omics information

3.1 Notations and settings

Various approaches have been proposed in the literature to predict survival times using high-dimensional data. Many of them can be seen as extensions or variants of the multivariate Cox regression model. Important exceptions include the nonparametric random forest procedure based on recursive partitioning [17, 18] or parametric and semi-parametric alternatives to the proportional hazard models [19].

In this paper, we deliberately consider only methods based on the Cox

model. The focus is on the construction of a linear predictor and on its ability to discriminate between good and bad prognosis patients rather than on the modelling of the relationship between linear predictor and survival time—although both aspects are of course tightly interconnected. Here we always assume a linear effect for all continuous variables and we consider only models with main effects.

From now on we thus assume that the hazard can be modeled as the product of a baseline hazard function $\lambda_0(t)$ with the exponential of a linear combination of the predictors:

$$\lambda(t|Z_1,\ldots,Z_q,X_1,\ldots,X_p) = \lambda_0(t)\cdot\exp(\gamma_1 Z_1+\cdots+\gamma_q Z_q+\beta_1 X_1+\cdots+\beta_p X_p),$$

where $\gamma_1,\ldots,\gamma_q$ and $\beta_1,\ldots,\beta_p$ are the regression coefficients of the clinical predictors $Z_1,\ldots,Z_q$ and of the omics predictors $X_1,\ldots,X_p$, respectively.

The goal of this paper is to investigate and propose procedures to estimate the regression coefficients $\gamma_1,\ldots,\gamma_q$ and $\beta_1,\ldots,\beta_p$, where the dimension q (number of clinical predictors) is typically small (say, from 1 to 10) and the dimension p (number of omics predictors) is typically very large (several hundreds or thousands). Because of the second—high-dimensional—part of the linear predictor $\gamma_1 Z_1+\cdots+\gamma_q Z_q+\beta_1 X_1+\cdots+\beta_p X_p$, the regression coefficients cannot be simply estimated as usual by maximization of the partial likelihood.

Thus, adaptations of the classical Cox regression have to be considered to address two embedded problems. The first problem is the handling of the high-dimensionality of the omics data that is challenging even in the absence of clinical predictors. This issue is addressed in Section 3.3. In the literature, high-dimensionality is typically handled via either variable selection, dimension reduction or regularization techniques. The second problem is the relative importance given to the clinical predictors and omics predictors respectively, an issue that we denote as the *combination* of low- and high-dimensional data and address in the Section 3.4. Different proposals have been made in the literature, often quite implicitly and without comparison to other approaches.

This paper aims to address the two issues outlined above simultaneously by presenting, illustrating and discussing the different associations of method handling high-dimensional predictors and combination scheme for accommodating low- and high-dimensional predictors into a single model. While all discussed methods handling high-dimensional data and combination schemes have been addressed previously in the literature, they have never been addressed in a unified framework, and many of them have never been considered together. This paper aims to fill this gap. More precisely, we suggest a general

framework formalizing the above mentioned issues and consider several pairs of methods handling high-dimensional predictors and combination schemes.

3.2 Handling low-dimensional clinical data

In the construction of clinical models we tried to include all the available information. For the breast cancer dataset, we included all the variables described in Table 1 of the original study [13], but *AJCC* for the aforementioned reason. Consistently with this approach, we used also the information on the number of lymph nodes involved (available on the repository although not reported/not used in the original study). We proceeded in the same way for the neuroblastoma data. Besides the information available in the dataset provided by Bøvelstad (i.e., *risk group*) we included the age at the diagnosis (available on the web repository). For some of the observations also the sex was available, but the number of missing values was too large, and therefore we decided to ignore this predictor in the clinical model. It is worth noting that the small predictors (we know only *age* and the *risk group*) may not cover all information generally available from the clinic and thus leaves more room for the added predictive value of the omics data. As model selection does not play a role in choosing the clinical model, the predictive ability should be similar in training and validation data.

3.3 Handling high-dimensional omics data

This section is concerned with the selection and estimation of effects from influential variables in high-dimensional data. As we consider survival time data in our two examples, methods will be discussed in the context of the Cox model. With the exception of the method with adjustment for clinical predictors in 3.3.5, we assume that only omics predictors are available.

3.3.1 Univariate variable selection (U)

The first—most naive—way to estimate a multivariate Cox model from high-dimensional data consists to consider the results of p univariate Cox models of the form

$$\lambda(t|X_j) = \lambda_0(t) \cdot \exp(\beta_j^{\mathrm{U}} X_j) \tag{1}$$

(for $j = 1, \ldots, p$) in order to select the most relevant variables, where the exponent 'U' in β_j^{U} stands for "univariate". In this paper, we fit a univariate Cox model for each considered predictor and then use the p-value of the likelihood ratio test as the measure of the predictor's relevance. To choose the number k of predictors to be selected, we consider multivariate Cox

models based on the k top predictors (according to the univariate p-value), with k varying from 1 to 25. The tuning parameter k is chosen via a 10-fold-cross-validation procedure by minimizing the sum of the integrated Brier score computed in each fold using the model trained in the other 9 folds.

Note that the choice of the number of predictors via cross-validation is not yet common practice in the literature—in contrast to the choice of, say, the penalty parameter in penalized regression, although these two types of parameters (penalty and number k of predictors) are of similar nature and should be handled similarly. By choosing k by cross-validation, we thus follow good statistical practice with respect to the important topic of parameter choice. We claim that in the literature k is too often chosen in an arbitrary way, possibly also hiding "fishing-for-significance" practices.

3.3.2 Forward variable selection (F)

The univariate variable selection approach outlined above has the major inconvenience that the set of predictors yielding the smallest p-values in univariate Cox regression may be very far from the optimal set of predictors in terms of prediction error, especially because of potentially strong correlations between them. To partially address this problem, an alternative is to use forward variable selection, a technique that is widely used in the context of low-dimensional data. In this paper we thus consider a forward selection procedure: starting from the null model, we add stepwise new predictors to the Cox model, using the p-value of the likelihood ratio test as entry criterion. We stop the procedure when the optimal number of predictors k is reached. Similarly to the univariate variable selection approach, k is computed via 10-fold-cross-validation by choosing the value, among the candidate $k = 1, \ldots, 25$, which minimize the integrated Brier score.

3.3.3 Lasso (L1)

The lasso technique [20] introduced by Tibshirani and adapted by the same author to survival analysis [21] is a penalized regression method, where the penalization term, based on the L_1 norm, forces many regression coefficients to be exactly 0 and thus allows to select a sparser model containing only the most relevant predictors, i.e. to perform intrinsic variable selection. This characteristic, together with the lasso's "shrinkage" property, has contributed to its large popularity in the context of high-dimensional omics data. The amount of penalization—and thus the sparsity of the resulting model—depends on the penalty parameter λ that is chosen via 10-fold-cross-validation in this paper. In order to have a fair penalty, the predictors are

scaled and forced to have variance 1. Since the method is based on the Cox model, centering the variables should not affect the estimate of the regression coefficient. In any case, here we choose to center the predictors around 0.

3.3.4 Boosting regression: offset boosting (Coxboost, CB) and gradient boosting (mboost, MB)

Boosting regression can be seen as another regularized regression technique that exploits the repeated fitting of a weak estimator in order to obtain step by step a good final model. The idea is to minimize stepwise a loss function: as common practice in survival analysis, the considered loss function is the negative partial log-likelihood or a penalized version of it. Here indeed we refer to two boosting techniques known as "offset boosting" [11] and "gradient boosting" [22]. The former is an adaptation of the boosting ridge regression [23] to survival analysis, and therefore the loss function is the negative partial log-likelihood with a L_2 penalization. The estimator is a first order approximation of the ridge estimator. The code to perform "offset boosting" in R is publicly available in the package CoxBoost [24]. With "gradient boosting", instead, we refer to the componentwise L_2-boosting technique [25, 26], and, in particular, to its version for survival data [17]. The loss function, in this case, is the negative partial log-likelihood, and the weak estimator is a weighted version of the ordinary least square estimator, repeatedly applied to the gradient of the partial log-likelihood [27] to obtain the final prediction model. The algorithm is implemented in R package 'mboost' [28].

Both boosting techniques depend on the two tuning parameters, namely a shrinkage factor and the number of boosting steps to perform. The former does not strongly affect the results as soon as it is set to a reasonable value. In this paper we use the procedures recommended by the respective authors: default value 0.10 for "gradient boosting", a coarse investigation for "offset boosting" (in our analysis, either as a result of the rough cross-validation-based routine *optimCoxBoostPenalty* or by considering a rough set of values in the selection of the number of boosting steps). Conversely, the latter is crucial and directly related to the complexity of the resulting final model. To set it, we use the 10-fold-cross-validated partial log-likelihood, available in the R implementation of the two boosting techniques. Finally, as recommended by the authors, we pre-process the data in order to be able to apply the aforementioned methods: for the "offset boosting", we center and scale the predictors, while for "gradient boosting" we limit to center them. Since it is not based on a penalty terms, indeed, the variance of the predictors is not relevant.

3.3.5 Adjustment to univariate (UA) and forward (FA) variable selection

The term "adjustment" refers to the adjustment for clinical predictors in the univariate and forward selection procedures, previously described, to take them into account when selecting omics predictors. The adjustment part is "somehow prespecified" (e.g., well established variables given from outside, determined in an earlier study or selected in a preliminary step) without considering the omics data. For adjustment of univariate selection (UA), in place of model (1), the model used to assess predictor X_j is then

$$\lambda(t|Z_1, \ldots, Z_q, X_j) = \lambda_0(t) \cdot \exp(\gamma_1^{\mathrm{UA}(j)} Z_1 + \cdots + \gamma_q^{\mathrm{UA}(j)} Z_q + \beta_j^{\mathrm{UA}} X_j),$$

where the exponent 'UA' in β_j^{UA} stands for "univariate with adjustment" and the exponent (j) indicates that these are the coefficients within the model assessing predictor X_j. In this way, the omics predictors are ranked based on their added predictive values to the clinical predictors and not to their predictive value itself. For example, a omics predictor highly correlated to the clinical data can be associated to a very small p-value in the simple univariate selection (and therefore added in the final model), but it is likely discarded within this adjusted procedure (larger p-value). Finally, the stopping criterion is determined by the number of predictors to be included, chosen by minimizing the integrated Brier score computed within a 10-fold-cross-validation procedure.

Adjustment for clinical predictors can also be performed within the forward selection procedure (FA). The only difference is then that omics predictors are added in a stepwise fashion by starting from the clinical model

$$\lambda(t|X) = \lambda_0(t) \cdot \exp(\gamma_1^{\mathrm{clin}} Z_1 + \cdots + \gamma_q^{\mathrm{clin}} Z_q) \tag{2}$$

instead of starting from the null model. The aim is again to select those omics predictors which better explain the outcome variability together and not independently to the clinical ones. Also in this case the number of predictors to consider in the model, again chosen by 10-fold-cross-validation minimizing of the integrated Brier score, represents the stopping criterion.

3.4 Combination of low- and high-dimensional data

No matter whether high-dimensional data are handled through e.g. univariate variable selection, lasso or boosting, it has to be defined whether clinical predictors and omics predictors should be treated differently, and if yes, how. This is what we denote as the combination strategy. Various existing general

strategies are outlined in this section and discussed in the specific context of clinical and omics predictors, following the line of Boulesteix and Sauerbrei [12].

3.4.1 Strategy 1: naive

The first strategy reviewed in the paper by Boulesteix and Sauerbrei [12] is called "naive" and consists of treating clinical and omics predictors in the same way. By definition this strategy cannot be applied together with the procedure adjusting for clinical predictors, since it ignores the difference between the two types of predictors. All predictors are merged together and no difference is done between clinical and omics predictors when applying univariate variable selection, forward variable selection, lasso regression or boosting regression (either CoxBoost or mboost), i.e. the clinical predictors are considered as X variables in the methods whose definition involves X_j.

The major inconvenience of this straightforward approach is that it ignores the information that clinical predictors are generally on average more predictive than omics predictors. Important clinical predictors are likely to be lost within omics predictors that look important as a result of multiple testing issues [11, 12].

3.4.2 Strategy 2: residuals

To address this issue, strategy 2 takes an opposite approach and first fully exploits the prediction potential of clinical predictors by fitting the clinical Cox model (2) to the data while completely ignoring omics predictors, yielding estimates $\hat{\gamma}_1^{\text{clin}}, \ldots, \hat{\gamma}_q^{\text{clin}}$ of the regression coefficients. Ideally it does not only consider linear predictors, but also assesses whether continuous variables have non-linear effects and it checks for potential interactions between predictors. However, here we restrict approaches to derive linear predictors. A method handling high-dimensional data such as those described in the previous section is then applied to the residuals of this model, i.e. by considering the fitted linear predictor

$$\hat{\eta}^{\text{clin}} = \hat{\gamma}_1^{\text{clin}} Z_1 + \cdots + \hat{\gamma}_q^{\text{clin}} Z_q$$

as an offset. Considering $\hat{\eta}^{\text{clin}}$ as an offset can be seen as equivalent to including η as a predictor in the model while forcing its coefficient to be 1. This strategy can be applied together with all considered methods handling high-dimensional data.

Together with the univariate variable selection method, including the linear predictor $\hat{\eta}^{\text{clin}}$ as an offset means that the linear predictor in the final

Cox model is estimated as

$$\hat{\eta}^{\text{clin}} + \sum_{j \in S} \hat{\beta}_j^{\text{U,offset}} X_j \;=\; \hat{\gamma}_1^{\text{clin}} Z_1 + \cdots + \hat{\gamma}_q^{\text{clin}} Z_q + \sum_{j \in S} \hat{\beta}_j^{\text{U,offset}} X_j$$

where the coefficient of $\hat{\eta}^{\text{clin}}$ is forced to be 1 and the coefficients $\beta_j^{\text{U,offset}}$ (for $j \in S$) are simply estimated by maximization of the partial likelihood. Here S is the set containing the indexes of the k omics predictors with smaller p-values. The final Cox models fitted for the forward variable selection can be expressed in a similar way by referring to its set of relevant predictors S. Moreover, we consider the "adjusted" versions of the univariate and forward selection within this strategy: although the estimates of γ_j^{clin} in $\hat{\eta}^{\text{clin}}$ are allowed to vary (this represents the main difference with the "not-adjusted" versions), we are selecting the omics predictors based on the residuals of the clinical model.

The 'residuals' strategy can also be adopted together with the lasso or both considered boosting techniques CoxBoost and mboost. With all these techniques, $\hat{\eta}^{\text{clin}}$ is entered in the model as an offset and the method are then applied as usual to the omics predictors. The final linear predictor that is then used to predict survival thus consists of two parts: the first part, $\hat{\eta}^{\text{clin}} = \hat{\gamma}_1^{\text{clin}} Z_1 + \cdots + \hat{\gamma}_q^{\text{clin}} Z_q$, is the same for all techniques, while the second part $\hat{\eta}^{\text{omic}}$ differs for the three techniques.

3.4.3 Strategy 3: "favoring"

The third strategy lies somehow between the previous two. It consists in favoring the clinical variables, in order to account for the known information that the clinical predictors have already shown their predictive value in the past. This subject-matter knowledge should be used (at least partly) in the model building process. In addition, it balances for the difference in cardinality between the clinical and omics sample spaces.

The univariate and forward variable selection procedures can be simply modified in order to increase the probability of selection of clinical predictors in an adequate way. The two types of predictors (clinical and omics) are considered as two separate blocks that are first analysed separately from each other. Univariate variable selection (resp. forward variable selection) is performed for each block separately—except for the final cross-validation procedure for the choice of the number of predictors to include in the model.

This final cross-validation procedure is conducted in a two-dimensional fashion, i.e. by evaluating in turn the value of the integrated Brier score obtained in the 10-fold cross-validation procedure for all pairs $(k_{\text{clin}}, k_{\text{omic}})$ of candidate numbers of predictors within a Cox model and by selecting the

pair which minimizes it. Here k_{clin} is allowed to vary between 0 and the total number q of clinical predictors, and k_{omic} is allowed to vary between 0 and 25. This two-dimensional scheme favors clinical predictors that have thus much higher chance to get included in the model than with the naive approach.

Finally, together with the univariate method, if S_{clin} and S_{omic} denote the set of the selected clinical and omics predictors, respectively, we obtain the linear predictor

$$\sum_{j \in S_{\text{clin}}} \hat{\gamma}_j^{U,favor} Z_j + \sum_{j \in S_{\text{omic}}} \hat{\beta}_j^{U,favor} X_j,$$

where the regression coefficients are estimated by maximization of the partial likelihood in the corresponding Cox model. For forward selection the linear predictor is similar, with S_{clin} and S_{omic} denoting the related sets of relevant predictors.

In order to adopt this "favoring" strategy via lasso, instead, we apply the penalty term only on the omics predictors, while the clinical variables are left un-penalized. The penalized partial log-likelihood is therefore in the form

$$pl_{pen}(\boldsymbol{\gamma}^{\text{L1}}, \boldsymbol{\beta}^{\text{L1}}, \lambda) = pl(\boldsymbol{\gamma}^{\text{L1}}, \boldsymbol{\beta}^{\text{L1}}) - \lambda \sum_{j=1}^{p} |\beta_j^{\text{L1}}|, \tag{3}$$

where $\boldsymbol{\gamma}^{\text{L1}}$ and $\boldsymbol{\beta}^{\text{L1}}$ stand for the vectors of regression coefficients for the clinical and omics predictors, respectively. The same idea works for the "offset boosting" technique: since it is based on the ridge regression, the negative partial log-likelihood (loss function) looks like Eq. (3) (obviously multiplied by -1), with the only difference on the penalty term, which is based on the L_2 norm and not the L_1 one. It is worth noting that this procedure forces to include the clinical predictors in the final Cox model, in contrast to the application of this strategy to univariate and forward selection.

3.4.4 Strategy 4a: Dimension reduction for omics predictors

Finally, the fourth strategy tackles the difference in dimensionality by summarizing the omics predictors into a single score that is then considered together with the clinical predictors in a Cox model. More precisely, as a first step we apply the chosen method for handling high-dimensional data to the omics predictors only. When the selection method directly provides a linear predictor, as in the case of lasso and the two boosting techniques, we use it as omics score. In the case of univariate variable selection, instead, we apply the principal component analysis on the space of the selected predictors, and we keep the first principal component (but, in principle, more can be used) as the score. This procedure is known in the literature as "supervised principal

	Strategies to combine clinical and omics data				
	1	2	3	4a	4b
selection method	naive	residuals	favoring	mol. score	scores
univariate selection (U)	✔	✔	✔	✔	✔
forward selection (F)	✔	✔	✔	✔	✔
lasso (L1)	✔	✔	✔	✔	✔
offset boosting (CB)	✔	✔	✔	✔	✔
gradient boosting (MB)	✔	✔	✘	✔	✔
univariate sel. with adj. (UA)	✘	✔	✘	✔	✔
forward sel. with adj. (FA)	✘	✔	✘	✔	✔

Table 1: All the combinations strategy/predictive method considered in this paper.

component" [29]. A similar procedure is followed for the adjusted version of univariate selection and for forward selection (with or without adjustment). Anyway, regardless of which technique we use to compute the score, this is finally used, together with the clinical variables, to obtain the final Cox model.

3.4.5 Strategy 4b: Dimension reduction for clinical predictors and omics predictors

To be even fairer in the dimensionality comparison, it is possible to summarize also the clinical information into one score. The easiest way to exploit this variation of strategy 4 is by using the linear predictor of the clinical Cox model as the clinical score. This score and the omics one computed as above are then used as explanatory variables in the final Cox model. This strategy is very similar to strategy 4a: the only difference is that in strategy 4a the coefficients of the clinical predictors are fit individually, while in strategy 4b they are re-fit together through the estimation of the coefficient of the clinical score as a whole.

3.5 Evaluation criteria

Table 1 summarizes all the possible approaches considered in this paper. The further step is to evaluate the prediction ability of the models resulting from these combinations. As a measure of the overall prediction ability, we take advantage of the integrated Brier score [30]. This measure, indeed, captures both the aspect of a good prediction, namely the calibration (similarity

between the actual and predicted outcomes) and the discrimination ability (ability of predicting the survival times of the observations in the right order) [11, 31]. The time-dependent Brier score is a quadratic score based on the predicted time-dependent survival probability that, ideally, should be 1 at time t if the subject i is alive and 0 otherwise [32] and takes censoring into account. The integrated Brier score is obtained by integrating the Brier score over the time t. It is worth noting that we take advantage of the integrated Brier score also in the 10-fold cross-validation procedure performed to select the number of relevant predictors in the univariate and forward selection (with and without adjustment). The integrated Brier score, indeed, is computed in each fold for the prediction models computed within the desired combination selection method/strategy with a number of predictors from 1 to 25, fitted in the remaining 9 folds. The requested value of the tuning parameter corresponds to the number of predictors of the model for which the sum of the 10 integrated Brier scores is minimum.

Since an application of a re-calibration procedure is always possible, it may also be advantageous to compare the models only in terms of discriminative ability. Usually this property is measured by estimating the concordance probability, i.e. the probability that two observations are correctly ranked by the model with respect to their survival time, through a suitable score. The most popular score used for this purpose is the C-index [33], which estimates the concordance probability by counting the proportion of the usable pairs that are concordant [33]. Here "usable" means that censoring does not prevent to order them, while "concordant" means that the actual and predicted survival times have the same ordering. As a ranked-order statistic, the C-index is totally insensitive to error in calibration.

These measures can be used to assess models both on the training data and the validation data. However, the results obtained on the former are usually much too optimistic and the different strategies cannot be compared depending on them [1, 34].

3.6 Implementation

We developed an R package, called *combPreds*, which contains all the functions useful to perform the analyses presented in this paper. To implement them, it takes advantage of existing packages, namely *survival* [35] (for Cox proportional hazard model), *glmnet* [36] (for lasso), *mboost* [28] (for gradient boosting), *CoxBoost* [24] (for offset-based boosting) and *pec* [37] (for computation of integrated Brier score and C-index). The package provides functions to automatically perform univariate and forward selections (with and without adjustment), together with a function useful to obtain the best

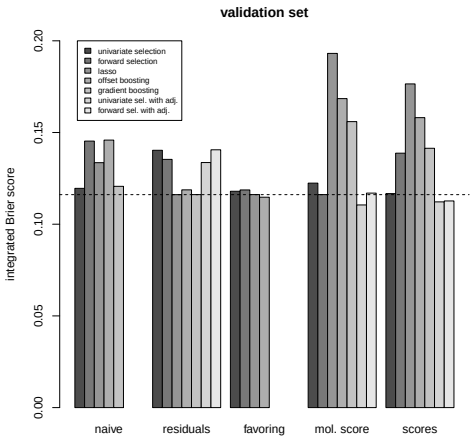

Figure 1: Breast cancer data: integrated Brier score, computed in the validation dataset up to 5 years, for all the possible combinations selection method/strategy. The dotted line represents the value for the clinical model.

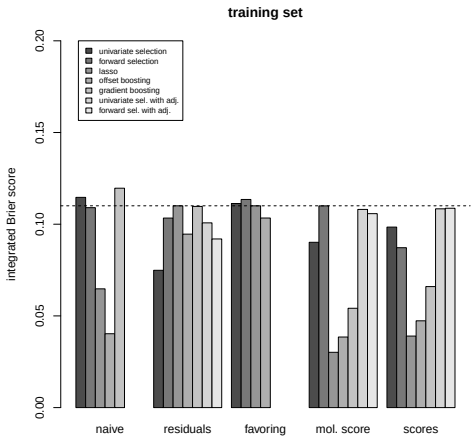

Figure 2: Breast cancer data: integrated Brier score, computed in the training dataset up to 5 years, for all the possible combinations selection method/strategy. The dotted line represents the value for the clinical model.

number of predictors (tuning parameter) in these four methods for each of the reviewed strategies. Other functions permit to compute the prediction error curves, the integrated Brier score and the C-index for all the statistical tools considered here. For all the other aspects (tuning parameter and model fitting in boosting and lasso), we used the functions from the original packages (*mboost*, *CoxBoost*, *glmnet*). The package *combPreds* and the R-code necessary to reproduce the analyses are available at *http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/de_bin/index.html*. The analysis performed on the breast cancer example is totally reproducible, while the analysis performed on the neuroblastoma example needs a dataset kindly provided by Hege Maria Bøvelstad, which is not directly accessible on the web.

Table 2: Breast cancer data: clinical model based on 283 observations (58 events). The column "coeff" reports the log-hazard-ratio with respect to the baseline, the column "distr" indicates the number of observations for each modality of the predictors (with percentage between brankets).

variable	distr	coeff	sd(coeff)	p-value
age<40	53 (0.19)	0.000	NA	NA
age≥ 40	230 (0.81)	-0.305	0.313	0.330
pr=-	151 (0.53)	0.000	NA	NA
pr=+	132 (0.47)	0.068	0.387	0.860
er=-	120 (0.42)	0.000	NA	NA
er=+	163 (0.58)	-0.896	0.415	0.031
t stage=01	21 (0.07)	0.000	NA	NA
t stage=2	155 (0.55)	0.356	0.621	0.567
t stage=3	60 (0.21)	0.484	0.663	0.465
t stage=4	47 (0.17)	0.874	0.659	0.185
nodal status=0	84 (0.30)	0.000	NA	NA
nodal status=1	131 (0.46)	1.402	0.489	0.004
nodal status=2	38 (0.13)	1.677	0.564	0.003
nodal status=3	30 (0.11)	1.904	0.559	0.001
grade=1	19 (0.07)	0.000	NA	NA
grade=2	115 (0.40)	1.025	0.982	0.326
grade=3	149 (0.53)	0.874	1.062	0.410

Table 3: Breast cancer data: number of clinical (first position) and omics (second position) predictors respectively selected. The symbol "*" indicates that the number of predictors is fixed by the method, the symbol † that the predictors are summarized in a single score.

	Strategies to combine clinical and omics data				
	1	2	3	4a	4b
selection method	naive	residuals	favoring	mol. score	scores
univariate selection (U)	$(0-2)$	$(6^{*\dagger}-10)$	$(4-0)$	$(6^*-13^\dagger)$	$(6^{*\dagger}-5^\dagger)$
forward selection (F)	$(0-2)$	$(6^{*\dagger}-1)$	$(2-0)$	$(6^*-8^\dagger)$	$(6^{*\dagger}-3^\dagger)$
lasso (L1)	$(0-46)$	$(6^{*\dagger}-0)$	(6^*-0)	$(6^*-49^\dagger)$	$(6^{*\dagger}-49^\dagger)$
offset boosting (CB)	$(0-64)$	$(6^{*\dagger}-8)$	(6^*-9)	$(6^*-38^\dagger)$	$(6^{*\dagger}-38^\dagger)$
gradient boosting (MB)	$(0-9)$	$(6^{*\dagger}-1)$	-	$(6^*-32^\dagger)$	$(6^{*\dagger}-32^\dagger)$
univariate sel. with adj. (UA)	-	(6^*-1)	-	$(6^*-12^\dagger)$	$(6^{*\dagger}-25^\dagger)$
forward sel. with adj. (FA)	-	(6^*-3)	-	$(6^*-14^\dagger)$	$(6^{*\dagger}-25^\dagger)$
clinical model	(6^*-0^*)				

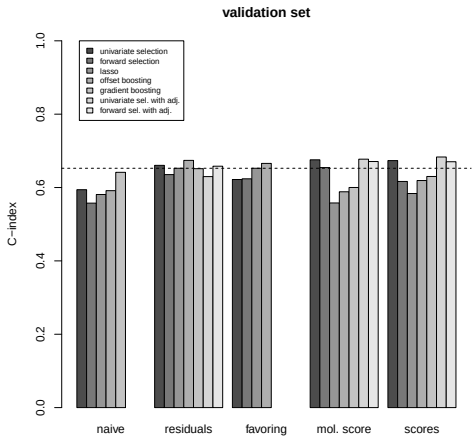

Figure 3: Breast cancer data: C-index for all the possible combinations selection method/strategy computed in the validation set. The dotted line represents the value for the clinical model.

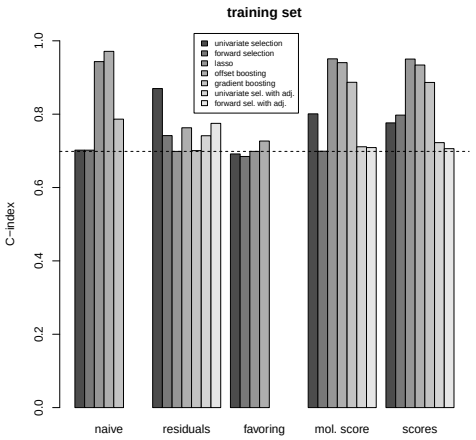

Figure 4: Breast cancer data: C-index for all the possible combinations selection method/strategy computed in the training set. The dotted line represents the value for the clinical model.

4 Results

4.1 Breast cancer data

The performances of the different combinations method/strategy on the breast cancer data in the validation set are reported in Figure 1 (integrated Brier score) and Figure 3 (C-index). The most obvious result is that almost all considered models are unable to noticeably exceed the predictive ability of the clinical model. It seems, therefore, that in this example the omics data have no added predictive value: in other words, their inclusion in a prediction model does not improve the predictive ability of the clinical model. This is not surprising, since in line with the results obtained by Bøvelstad and colleagues [10] in a different breast cancer dataset [38]. Furthermore, this example clearly illustrates the overfitting issues affecting prediction models based on high-dimensional data. It can be seen by comparing the integrated Brier score computed in the validation (Figure 1) and in the training (Figure 2) sets: in general, the more the model explains the outcome of the training set, the worse its predictive performance is in the validation set. In this dataset, the risk of facing the overfitting issue seems to be more relevant when we exploit the "naive" strategy: using a multivariate selection method (i.e.,

gradient boosting, offset boosting, lasso or forward selection) we see that the predictive performance of the prediction model fitted within this strategy is worse than those built within the "residuals" and "favoring" ones.

With regards to the overfitting issue, further hints can be seen in the results within strategies 4a and 4b. Let us firstly focus on those models for which the omics score is computed using the linear predictor of a model fitted to the omics data (offset boosting, gradient boosting, and lasso): the results in terms of both integrated Brier score and C-index are worse than the simple clinical model. This result can be interpreted as a combination of two mechanisms: the aforementioned overfitting to the training data, and the fact that, within this strategy, clinical and omics information are treated in a completely separate way. The omics score, indeed, is constructed in order to explain the outcome variability without taking into consideration the information provided by the clinical data, and the risk that both omics score and clinical data provide the same information is high. This latter issue seems to be confirmed by the fact that summarizing also the clinical information as a unique score (i.e., exploiting strategy 4b), the prediction ability of the models increases. This seems counter-intuitive, since in this way we are losing information, but it can make sense if we think that this information may be redundant (clinical predictors and omics score explain the same part of outcome variability) and may contribute to overfitting.

Within strategies 4a and 4b, methods that decrease the dependence of the omics score on the training set seem to yield better results. This is the case of univariate and forward selection, with and without adjustment, in which the first principal component is used as omics score. The principal component analysis, indeed, seems to be able to mitigate the effect of overfitting, even if, like in the case of the "adjusted" versions within strategy 4b, the number of predictors involved in the omics score is large (25, see Table 3). Conversely, it can be dangerous to construct a principal component with not enough predictors (forward selection/strategy 4b). Besides these considerations about the number of predictors involved, we point out that the performance of a model fitted within strategies 4a and 4b depends also on the amount of variance explained by the first principal component itself and its association with the outcome: if it is too large, overfitting may lead to important problems, but if it is too small, the predictive ability of the resulting omics score is also small. Unfortunately, it is difficult to handle this problem: an option would be to include several principal components instead of only one, but a model with more than one principal component makes the medical interpretation more difficult, and the choice of the number of principal components to be included would be a non-trivial issue. Despite this, the good results obtained using this strategy are in line with the findings of

those studies which introduce gene-expression signatures based on the first principal component of a suitably generated space of relevant omics predictors [39, 40]. It is worth noting that, while the cross-validation procedure for univariate and forward selection (with and without adjustment) optimizes the tuning parameter (number of relevant predictors) within the strategies, i.e. for the combined model, for CoxBoost, mboost and lasso it optimizes the tuning parameter (penalty term, number of boosting steps) to construct the omics score only.

Besides these considerations, the comparison between Figure 1 and Figure 2 highlights the importance of a validation in an independent dataset. This is also true when dealing with low-dimensional problems but becomes absolutely necessary when the prediction problem involves high-dimensional data. We can see that, for the clinical model, the integrated Brier score computed on the training data is not severely larger than the one computed in the validation set (0.110 versus 0.116, dotted lines), while the difference (over-optimism) can be huge for combined models.

Another fact that seems to be highlighted by the analysis of this dataset, and that is connected with the overfitting issue, is the advantage of fitting a "sparse" model. In this dataset, indeed, the more predictors we select, the higher is the chance to face overfitting. We have already stated that, within the 'naive" strategy, models derived performing offset boosting and lasso do not well perform. If we look at Table 3, we see that the number of predictors involved is high (64 and 46, respectively), while models with less predictors (e.g., 9 for the model fitted with a gradient boosting technique) perform better. This is even more important if we select the model within a strategy, like univariate selection and forward selection, which has no property that permits to attenuate the overfitting issue, for example shrinkage. It is probably for this reason that the cross-validation procedure, in these two cases, selects only 2 predictors for the best model.

Finally, we have already stated that in this dataset the clinical model (Table 2) has itself a prediction performance in line with the best combined models. This is confirmed by the performance of lasso and gradient boosting within strategy 2 and 3. Looking at the number of selected predictors reported in Table 3, in particular, we can see that they both yield the clinical model, i.e. select no omics predictor (lasso) or a model including only one omics predictor (gradient boosting). Since in these two methods the "favoring" approach consists in considering the clinical predictors as mandatory, the results for strategies 2 and 3 do not substantially differ. In any case, the only (really small) improvement to the clinical model is obtained within strategy 3 and offset boosting. This result is also in agreement with the findings of Bøvelstad and colleagues [10]: in their application to breast

Table 4: Neuroblastoma data: clinical model based on 240 observations (51 events). The column "coeff" refers to the log-hazard-ratio with respect to the baseline, the column "distr" indicates the number of observations for each modality of the predictors.

variable	distr	coeff	sd(coeff)	p-value
risk = low/intermediate	150 (0.62)	0.000	NA	NA
risk = high	90 (0.38)	2.855	0.424	1×10^{-11}
age < 444.5	95 (0.40)	0.000	NA	NA
age ≥ 444.5	145 (0.60)	-0.530	0.320	0.097

cancer data, indeed, the only method able to outperform the clinical model is ridge regression, which is the basis of this particular kind of boosting procedure. A more noticeable improvement by using the "favoring" strategy rather than the "residuals" strategy can be observed for the univariate and forward selection methods: here the "favoring" does not force the selection method to select all the clinical predictors and the univariate selection procedure selects only the relevant clinical predictors. The resulting model leads to better prediction ability, both with respect to the "naive" and "residuals" strategy. In particular, with respect to the latter, this result highlights the importance to have a good clinical model when constructing a combining prediction model, especially within strategy 2: this is an interesting issue, which needs more investigation. We will sketch some possible solutions in the Discussion section.

4.2 Neuroblastoma data

The considerations on this second dataset begin with the fact that, in this case, the prediction models derived by the different combinations strategy/method generally perform better than the clinical model (especially in discriminative ability, see Figure 6), although the differences are not huge. The clinical model is displayed in Table 4: it contains a strong predictor, namely the *risk group*, and one whose significance is border line, the *age* at the diagnosis. The regression coefficient of the former is very significant, and in this case it may happen that it is included in a prediction model also within the "naive" strategy (e.g., in this example, for gradient boosting).

In this dataset we can observe again some situations seen in the previous example. First of all, there seems to be a general (although small) advantage of following "residuals" and "favoring" strategies rather than the "naive" one in the construction of a predictive model based on the boosting and lasso techniques. A behaviour similar to the breast cancer example occurs

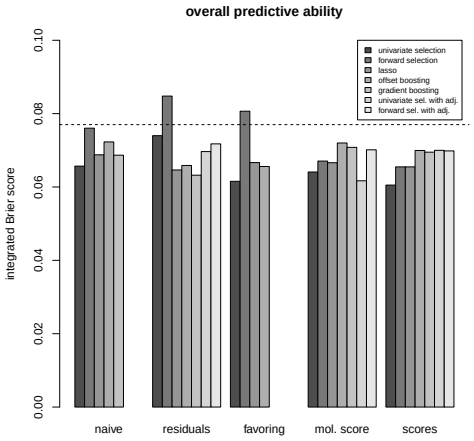

Figure 5: Neuroblastoma data: integrated Brier score, computed in the validation dataset up to 5 years, for all the possible combinations selection method/strategy. The dotted line represents the value for the clinical model.

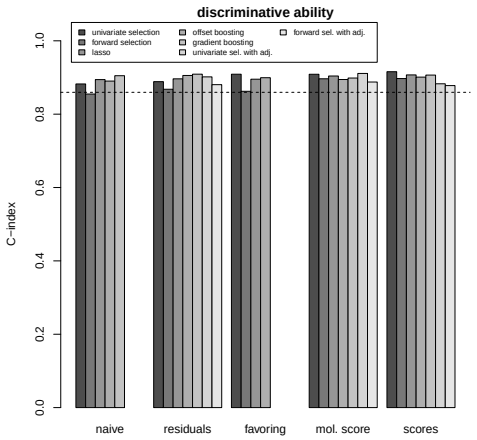

Figure 6: Neuroblastoma data: C-index for all the possible combinations selection method/strategy in the validation data. The dotted line represents the value for the clinical model.

Table 5: Neuroblastoma data: number of clinical (first position) and omics (second position) predictors considered. The symbol "*" indicates that the number of predictors is fixed by the method, the symbol † that the predictors are summarized in a single score.

	Strategies				
	1	2	3	4a	4b
Selection method	naive	residuals	favoring	mol. score	scores
univariate selection (U)	$(0-5)$	$(2^{*\dagger}-21)$	$(1-5)$	$(2^{*}-24^{\dagger})$	$(2^{*\dagger}-25^{\dagger})$
forward selection (F)	$(0-1)$	$(2^{*\dagger}-1)$	$(2-3)$	$(2^{*}-16^{\dagger})$	$(2^{*\dagger}-16^{\dagger})$
lasso (L1)	$(0-26)$	$(2^{*\dagger}-5)$	$(2^{*}-19)$	$(2^{*}-29^{\dagger})$	$(2^{*\dagger}-29^{\dagger})$
offset boosting (CB)	$(0-17)$	$(2^{*\dagger}-8)$	$(2^{*}-9)$	$(2^{*}-19^{\dagger})$	$(2^{*\dagger}-19^{\dagger})$
gradient boosting (MB)	$(1-17)$	$(2^{*\dagger}-17)$	-	$(2^{*}-18^{\dagger})$	$(2^{*\dagger}-18^{\dagger})$
univariate sel. with adj. (UA)	-	$(2^{*}\dagger-16)$	-	$(2^{*}-23^{\dagger})$	$(2^{*\dagger}-22^{\dagger})$
forward sel. with adj. (FA)	-	$(2^{*}\dagger-19)$	-	$(2^{*}-25^{\dagger})$	$(2^{*\dagger}-25^{\dagger})$
clinical model	$(2^{*}-0^{*})$				

also for the univariate selection method: the approach based on the "residuals" strategy performs quite badly, probably due to the too large number of selected variables (21, see Table 5), which may cause overfitting to the training data. On the contrary, a model constructed with this method within the "favoring" strategy works relatively well, leading to results which are among the best in this example. Furthermore, we see also in this example the possible advantage of using the first principal component instead of a linear predictor in constructing the omics score for strategies 4a and 4b. In particular, again the models constructed with the univariate selection with adjustment have very good performances both in terms of overall prediction ability (integrated Brier score, Figure 5) and discriminative ability (C-index, Figure 6). In this dataset, the results for strategies 4a and 4b do not substantially differ from each other; this can probably be explained by the small amount of clinical predictors: the results do not noticeably change whether we summarize the clinical information in one score or not.

In this example, it seems that the omics predictors can add predictive value to the clinical data in the fitting of a prediction model. It is worth noting that the simple forward selection is not able to capture this added value: the cross-validation procedure, indeed, selects only few omics predictors for strategies "naive", "residuals" and "favoring". This is maybe due to the fact that, without any property attenuating the overfitting issue (conversely to boosting and lasso techniques), this method cannot manage enough omics predictors without leading to a model strongly related with the training data. This is supported by the results for strategies 4a and 4b: the reduction of the overfitting issue obtained by the principal component analysis (performed to construct the omics predictor) permits the method to handle more omics predictors (in this case 16, see Table 5), leading to a prediction model with good predictive and discriminative ability.

5 Discussion

We have demonstrated how low-dimensional clinical data and high-dimensional omics data can be combined into a global prediction model incorporating the two types of information. The high dimension of the omics data makes most classical approaches inapplicable and leads to substantial problems such as overfitting or the risk to not fully exploit relevant information from the clinical data.

As far as the different investigated combinations between strategy and selection methods are concerned, none of them outperforms the others in both datasets. In this context it is important to note that:

- This is something expected, since each selection method and each strategy has its own advantages, which can be more or less relevant depending on the specific characteristics of a dataset. We think, therefore, that it is useful to have provided several options that researchers can exploit in their analysis.

- At the same time, it could be interesting to better investigate which are those characteristics of a dataset which correspond to a better performance of a specific combination selection method/strategy. This can be done in a systematic simulation study, which is in our plans.

- Even if our study had identified a clear winner over the two investigated datasets, it would not necessarily be representative of the performance on the whole domain of interest. Making general conclusions on the performance of prediction approaches based on only $n = 2$ datasets would be as if a medical doctor would make conclusions on the efficacy of treatments based on only $n = 2$ patients: nonsense! Hence, our study is definitely meant as illustrative. Even if we try to interpret the results, it does not mean that this interpretation is universally valid and that similar results would be observed for other similar datasets. A thorough discussion of these problems can be found elsewhere [41].

- Even if there were a real winner method that truly performs best on average over the datasets of interest, the power of our study with $n = 2$ datasets would be too limited to discover it [42]. We again point out that our comparison study is of illustrative nature.

That said, our study illustrates important aspects that have to be taken into account when fitting a combined model. Firstly, the combined use of clinical and omics predictors makes sense only when the latter contain added predictive value, and do not simply provide similar information as the clinical ones. This point is highly relevant: we have seen in our first example how the clinical model competes in terms of predictive ability with all the other models, which are built by taking advantage also of the omics information.

Secondly and in the same vein, it is important to have a good clinical model. The presence of an irrelevant predictor can worsen the results. The analysis of the breast cancer dataset shows that this problem can affect the prediction ability of those models built within the "residual" strategy or those implementations of the "favoring" strategy which consider the clinical predictors as mandatory (namely lasso and boosting). Even if the effect of including an irrelevant predictor does not directly worsen the prediction ability, it affects some other properties of a prediction model, such as its

sparsity and, somehow correlated, the simplicity of its interpretation. An issue related to the adequateness of the clinical model is whether the effects are all additive—as implicitly assumed in our paper— or whether, e.g., non-linear effects or interactions may be present. A too simplistic clinical model may artificially increase the apparent added predictive value of omics data, especially if clinical and omics data are correlated.

A possible solution to these problems in the context of the residuals strategy is to exploit a model selection procedure (for example, based on AIC) or a procedure for handling non-linear effects in order to use a better linear predictor as an offset. This solution is practicable only when the following step does not involve the clinical predictors ("residuals" and "dimensionality reduction"), and therefore cannot be easily extended to the favoring strategy. It is worth noting that also the two selection methods with adjustment for the clinical predictors do use their information in the building procedure, and therefore in these cases a different solution should be found, in order to not use twice the clinical data (firstly to select the best clinical model, then to fit the prediction model). It is worth noting, however, that a preselection of the clinical model may introduce some bias due to the model selection procedure, and the inference/interpretation of the regression coefficients may be incorrect [43].

Finally, another critical topic that we have not deeply considered here is the choice of the tuning parameter, which may be problematic and strongly affect the performance of the selection methods. In our study we tried to be as fair as possible by implementing the same procedure (maximization of a cross-validated likelihood based on 10 folds split) for all the selection methods, but this aspect should be investigated more deeply. In particular, the randomness of the split into 10 folds can lead to results that may be slightly different, especially in the case of high-dimensional data. This introduces some variance in the results, which should be taken into consideration when comparing the different strategies. However, this issue is more relevant in terms of identification of the significant variables than to the overall predictive performance itself. In any case, possible solutions to partly reduce the variability due to cross-validation are available in the literature, based on bootstrap [44] or repetition of the cross-validation procedure [45, 46] at the expense of computational time.

References

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

[2] Dobbin KK, Beer DG, Meyerson M, Yeatman TJ, Gerald WL, Jacobson JW, Conley B, Buetow KH,

Heiskanen M, Simon RM, *et al.*. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clinical Cancer Research* 2005; **11**:565–572.

[3] Ioannidis JP. Microarrays and molecular research: noise discovery? *The Lancet* 2005; **365**:454–455.

[4] Ioannidis J. Expectations, validity, and reality in omics. *Journal of Clinical Epidemiology* 2010; **63**:945–949.

[5] Hess K, Wei C, Qi Y, Iwamoto T, Symmans WF, Pusztai L. Lack of sufficiently strong informative features limits the potential of gene expression analysis as predictive tool for many clinical classification problems. *BMC Bioinformatics* 2011; **12**:463.

[6] Nevins JR, Huang ES, Dressman H, Pittman J, Huang AT, West M. Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Human Molecular Genetics* 2003; **12**:R153–R157.

[7] Stephenson AJ, Smith A, Kattan MW, Satagopan J, Reuter VE, Scardino PT, Gerald WL. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer* 2005; **104**:290–298.

[8] Obulkasim A, Meijer G, van de Wiel M. Stepwise classification of cancer samples using clinical and molecular data. *BMC Bioinformatics* 2011; **12**:422.

[9] Ideker T, Dutkowski J, Hood L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell* 2011; **144**:860–863.

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

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

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

[13] Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, *et al.*. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *Journal of the American Medical Association* 2011; **305**:1873–1881.

[14] Oberthuer A, Kaderali L, Kahlert Y, Hero B, Westermann F, Berthold F, Brors B, Eils R, Fischer M. Subclassification and individual survival time prediction from gene expression data of neuroblastoma patients by using caspar. *Clinical Cancer Research* 2008; **14**:6590–6601.

[15] Hudis CA, Barlow WE, Costantino JP, Gray RJ, Pritchard KI, Chapman JAW, Sparano JA, Hunsberger S, Enos RA, Gelber RD, *et al.*. Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the steep system. *Journal of Clinical Oncology* 2007; **25**:2127–2132.

[16] Sauerbrei W, Royston P, Bojar H, Schmoor C, Schumacher M. Modelling the effects of standard prognostic factors in node-positive breast cancer. *British Journal of Cancer* 1999; **79**:1752.

[17] Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ. Survival ensembles. *Biostatistics* 2006; **7**:355–373.

[18] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *The Annals of Applied Statistics* 2008; :841–860.

[19] Schmid M, Hothorn T. Flexible boosting of accelerated failure time models. *BMC Bioinformatics* 2008; **9**:269.

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

[21] Tibshirani R. The lasso method for variable selection in the Cox model. *Statistics in Medicine* 1997; **16**:385–395.

[22] Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* 2001; **29**:1189–1232.

[23] Tutz G, Binder H. Boosting ridge regression. *Computational Statistics & Data Analysis* 2007; **51**:6044–6059.

[24] Binder H. *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks* 2011. URL `http://CRAN.R-project.org/package=CoxBoost`, R package version 1.3.

[25] Bühlmann P, Yu B. Boosting with the L 2 loss: regression and classification. *Journal of the American Statistical Association* 2003; **98**:324–339.

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

[27] Ridgeway G. Generalization of boosting algorithms and applications of bayesian inference for massive datasets. PhD Thesis, University of Washington 1999.

[28] Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. *CoxBoost: Cox models by likelihood based boosting for a single survival endpoint or competing risks* 2013. URL `http://CRAN.R-project.org/package=mboost`, R package version 2.2-2.

[29] Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology* 2004; **2**:e108.

[30] Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* 1999; **18**:2529–2545.

[31] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, Pencina MJ, Kattan MW. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology* 2010; **21**:128.

[32] Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. *Bioinformatics* 2007; **23**:1768–1774.

[33] Harrell F, Lee KL, Mark DB. Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 1996; **15**:361–387.

[34] Castaldi PJ, Dahabreh IJ, Ioannidis JP. An empirical assessment of validation practices for molecular classifiers. *Briefings in Bioinformatics* 2011; **12**:189–202.

[35] T T. *A Package for Survival Analysis in S* 2013. URL `http://CRAN.R-project.org/package=survival`, R package version 2.37-4.

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

[37] Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *Journal of Statistical Software* 2012; **50**:1–23.

[38] van Houwelingen HC, Bruinsma T, Hart AA, van't Veer LJ, Wessels LF. Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine* 2006; **25**:3201–3216.

[39] Metzeler KH, Hummel M, Bloomfield CD, Spiekermann K, Braess J, Sauerland MC, Heinecke A, Radmacher M, Marcucci G, Whitman SP, *et al.*. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* 2008; **112**:4193–4201.

[40] Herold T, Jurinovic V, Metzeler K, Boulesteix AL, Bergmann M, Seiler T, Mulaw M, Thoene S, Dufour A, Pasalic Z, *et al.*. An eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia. *Leukemia* 2011; **25**:1639–1645.

[41] Boulesteix AL. On representative and illustrative comparisons with real data in bioinformatics: response to the letter to the editor by smith et al. *Bioinformatics* 2013; **29**:2664–2666.

[42] Boulesteix AL, Hable R, Lauer S, Eugster M. A statistical framework for hypothesis testing in real data comparison studies. *Technical Report 136*, University of Munich 2013. URL `http://epub.ub.uni-muenchen.de/14324/1/TR.pdf`.

[43] Leeb H, Pötscher BM. Model selection and inference: Facts and fiction. *Econometric Theory* 2005; **21**:21–59.

[44] Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 1992; **11**:2093–2109.

[45] Martinez JG, Carroll RJ, Müller S, Sampson JN, Chatterjee N. Empirical performance of cross-validation with oracle methods in a genomics context. *The American Statistician* 2011; **65**:223–228.

[46] Boulesteix AL, Richter A, Bernau C. Complexity selection with cross-validation for lasso and sparse partial least squares using high-dimensional data. *Algorithms from and for Nature and Life*. Springer, 2013; 261–268.