LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Anne-Laure Boulesteix, Willi Sauerbrei

Added predictive value of high-throughput molecular data to clinical data, and its validation

Technical Report Number 087, 2010
Department of Statistics
University of Munich

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

Added predictive value of high-throughput molecular data to clinical data, and its validation

Anne-Laure Boulesteix[1] and Willi Sauerbrei[2]

September 17, 2010

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany

[2] Department of Medical Biometry and Medical Informatics, Universitätsklinikum Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany

Corresponding author:

Anne-Laure Boulesteix
Tel.: +49 89 7095-7598
Fax: +49 89 7095-7491
Email: boulesteix@ibe.med.uni-muenchen.de

Keywords: Validation, added predictive value, clinical usefulness, independent data, prediction models, survival analysis, supervised classification

Abstract

Hundreds of "molecular signatures" have been proposed in the literature to predict patient outcome in clinical settings from high-dimensional data, many of which eventually failed to get validated. Validation of such molecular research findings is thus becoming an increasingly important branch of clinical bioinformatics. Moreover, in practice well-known clinical predictors are often already available. From a statistical and bioinformatics point of view, poor attention has been given to the evaluation of the added predictive value of a molecular signature given that clinical predictors are available. This article reviews procedures that assess and validate the added predictive value of high-dimensional molecular data. It critically surveys various approaches for the construction of combined prediction models using both clinical and molecular data, for validating added predictive value based on independent data, and for assessing added predictive value using a single data set.

1 Introduction

While high-throughput molecular data such as microarray gene expression data have been used for disease outcome prediction or diagnosis purposes for more than ten years [1] in biomedical research, the question of the added predictive value of such data given that classical clinical predictors are already available has long been under-considered in the bioinformatics literature.

This issue can be summarized as follows. For a given prediction problem (for example tumor subtype diagnosis or long-term outcome prediction), two types of predictors are considered. On the one hand, conventional clinical predictors such as, e.g. age, sex, disease duration or tumor stage are available as potential predictors. They have often been extensively investigated and validated in previous studies. On the other hand, we have molecular predictors which are generally much more difficult to measure and collect than conventional clinical predictors, and not yet well-established. In the context of translational biomedical research, investigators may be interested in the added predictive value of such predictors over classical clinical predictors. Note that clinical predictors may be given as a list of individual factors or in form of a well-established index such as the IPI index for lymphoma [2] or the Nottingham prognostic index for breast cancer [3]. In this paper we do not distinguish between the case of individual clinical predictors and the case of an aggregated index: from a statistical point of view, an aggregated index can be seen as a clinical predictor.

A particular challenge when assessing the added predictive value of molecular data is that these data are often high-dimensional, with typically $p \propto 10,000$ candidate predictors in the special case of gene expression microarrays. This potentially leads to overfitting problems and overoptimistic conclusions on their added predictive power [4, 5] when researchers work with a single data set at hand. As a result of high dimension, it is almost always possible to find a combination of molecular predictors that are associated with the outcome in the considered data set, independently of the true predictive power. Thus, building a molecular score based on the available data set and then testing its significance in multivariate analysis while adjusting for clinical predictors is not appropriate. It often yields a dramatic over-estimation of the molecular predictors' relevance: because the score is derived by "fishing" for relevant predictors within a huge number of molecular

predictors, it considerably overfits the data at hand. While this problem essentially affects all data analyses, it is strongly amplified in high-dimensional settings.

Validation of prediction models using independent validation data is a crucial step that is always necessary before clinical applications [6, 7, 8, 9] and now required by many high-ranking journals. See the paper by Castaldi and colleagues in this special issue for a survey of concrete studies including a validation step. By validation of a prediction model, researchers often mean the evaluation of the prediction model's error when predicting new observations from the validation data set, or some test of association between the derived score and the outcome of interest based on the validation data.

Going one step further, George [8] states that "the purpose of validation is not to see if the model under study is "correct" but to verify that it is useful, that it can be used as advertised, and that it is fit for purpose". To verify that the model is useful, validation of the predictive ability of the model is not sufficient as the clinical interest centers around the added value compared to previous existing models [10]. To verify that the new model is useful, one also needs to *validate the added predictive value*. This concept is not trivial from a methodological point of view and one may think of many different procedures in this context. This paper discusses statistical techniques and gives some recommendations for practical applications to high-throughput data.

Note that Altman and Royston [11] distinguish "statistical validity" from "clinical validity", which is another important aspect of validation of prognostic models. The latter issue is related to model stability, simplicity and transportability. These aspects are important for external validity [12] but beyond the scope of this paper. The focus of this review is on statistical validity.

From now on, we consider that two data sets are available to the researchers. The *training data set* is available from the beginning of the project and used for various statistical analyses, for instance for deriving a prediction model or a score (see Section 2 for the definition of prediction models and scores). The *validation data set* is used to assess and *validate* the results of the training phase. Ideally, it is not even opened until the end of the training phase to avoid "optimal selection" mechanisms [13]. In practice, validation data sets are often data collected later (temporal validation) or data collected elsewhere (external validation) [11]. But the training and test sets may also be drawn randomly from a single (large) data set at hand. already from traditional research the necessity of external validation is well-known [14, 15]: it is the only way to ensure that the derived model is widely useful and does not only work in the particular setting in which it was developed.

The rest of the paper is structured as follows. Section 2 introduces the terminology of prediction models used in this paper. Section 3 gives an overview of possible methods for deriving combined prediction models based on both clinical and molecular predictors. Section 4 presents several existing approaches to validate added predictive value based on training and validation data, while Section 5 briefly reviews procedures based on a single data set. Before the concluding remarks we will discuss further evaluation procedures in Section 6.

2 Prediction models

In this paper, we focus on two important prediction problems encountered in biomedical applications: *binary class prediction* and *prediction of survival*. The outcome is a

binary class label in the case of binary class prediction, for instance responder versus non-responder or healthy versus diseased. In survival analysis, the outcome is a right-censored time-to-event such as the time to death or the time to next relapse. We will consider logistic regression (for binary class prediction) and Cox regression (for survival analysis) as standard multivariate methods for data with much less independent variables than observations, although alternative approaches are conceivable like, e.g. accelerated failure time models (for survival analysis) or probit regression (for binary class prediction).

The molecular predictors measured through high-throughput experiments (like microarrays) are denoted as $X_1, \ldots, X_p$, where p is possibly as large as several tens of thousands and most often exceeds the sample size n. Classical clinical predictors such as age, sex, or TNM status are denoted as $Z_1, \ldots, Z_q$ with q commonly ranging from one to about fifteen. Whereas the molecular predictors $X_1, \ldots, X_p$ are measured at the same scale (most often a metric scale), the clinical variables $Z_1, \ldots, Z_q$ may be categorical (e.g. sex, TNM, ER status, mutational status), metric (e.g. tumor size, age), or a metric variable categorized by using one or more cutpoints.

In our context, a *prediction model* is defined as a function that assigns a class (in the case of class prediction) or a survival function estimate (in the case of survival analysis) to each new observation. Note that many class prediction methods also output estimated probabilities for each class in addition to the predicted class. In this paper, the term *score* denotes an index computed based on a number of candidate predictors that is supposed to be associated with the outcome of interest. Linear scores are a most important example in practice. For instance, a 3-genes linear score may be given as

$$\text{score} \;=\; -0.113 \times \text{geneA} \;+\; 0.207 \times \text{geneB} \;+\; 0.091 \times \text{geneC}, \qquad (1)$$

where "geneA", "geneB" and "geneC" stand for the respective expression levels of these genes. Prediction models are most often based on such scores, but a score is not sufficient to specify a prediction model. Linear scores can be derived by, e.g. lasso regression [16], elastic nets [17, 18], superPC [19], or Cox regression performed after univariate filtering. Clinical scores are usually derived with standard variable selection methods based on logistic regression or Cox regression. Widely speaking, estimated class probabilities returned by ensemble methods like random forests [20] in the case of class prediction can also be considered as (non-linear) scores. Note that in this case the score does not have a simple closed form like Eq.(1), and that the score is actually the prediction model itself. In general, however, the score does not fully specify the prediction model. In generalized linear models the estimated intercept and the link function are needed in addition to the linear score of the form (1). For class prediction one or more cutpoints are needed to separate patients into several groups. In Cox regression the prediction model consists of the combination of the score with the estimated cumulative hazard function. Prediction models and scores may involve only clinical predictors (*clinical prediction model/score*), only molecular predictors like in the example above (*molecular prediction model/score*) or a combination of both. These definitions are summarized in Table 1.

3 Strategies to derive combined prediction models

Prediction models combining clinical with molecular data are important to assess the added predictive value of molecular predictors. That is because some methods for assessing added predictive value are based on the comparison of the accuracy of prediction models with and without molecular predictors. However, the concept of combined models is not clearly defined and different strategies have been adopted in the literature. The important characteristics of the five strategies outlined below are summarized in Table 2.

3.1 Strategy 1 ("naive")

The perhaps most natural approach consists in building a combined prediction model by treating clinical and molecular predictors in the same way. This approach is very general: it can be applied to any prediction method that can handle predictors of the considered types (for instance a mixture of continuous molecular predictors and categorical clinical predictors). In this approach, individual clinical predictors may "get lost" within the numerous molecular predictors and thus not be fully exploited – especially when clinical information is available in form of a single aggregated score. If the clinical predictors have good predictive value, such naive prediction models are expected to underestimate the accuracy of combined models. The estimated added predictive value then tends to be small – not because the molecular predictors are bad but because the combined rule does not fully exploit the clinical predictors (that are lost within a large amount of noise).

3.2 Strategy 2 ("residuals")

The other extreme strategy consists in deriving a fixed clinical prediction model, for instance using logistic regression or Cox regression. The resulting linear predictor is then considered as an offset and updated using molecular predictors, for instance via lasso regression [16] or boosting regression [21]. This approach yields a linear predictor in which the coefficients of the clinical variables are prone to selection bias [12] but are not affected by the molecular predictors. It is adequate to test added predictive value [22]: the focus is here on the residual variation of the outcome. However, it may be sub-optimal in terms of prediction accuracy: depending on the correlation between clinical and molecular predictors, accuracy may be improved by adapting the coefficients of clinical predictors [23].

An important variant of this strategy is when a clinical model is already given, e.g. as an established index from the literature. Strategy 2 can also be applied in this case: the only difference is that the clinical model is not estimated from the data. This avoids a potential bias caused by building the clinical model, but in principle it does not change the way in which the molecular component of the combined score is derived.

Another variant of this strategy where the clinical component of the combined prediction model is not affected by molecular predictors consists in defining subgroups based on the clinical predictors and then fitting molecular prediction models in each subgroup separately. In this setup, interaction effects between clinical and molecular predictors may lead to substantially different prediction models in the considered subgroups. Simple examples may be separate investigations in groups defined by sex or by menopausal status in women. However, sample sizes are often (much) too small for such investigations,

especially if there are several important clinical predictors.

3.3 Strategy 3 ("favoring")

An intermediate strategy between strategies 1 and 2 is to fit a prediction model to clinical and molecular predictors simultaneously while somehow "favoring" clinical predictors, since they are more or less "established" prognostic factors. For instance, this can be done in terms of prior in Bayesian settings or through a different penalty in penalized regression. For example, the R package **penalized** [24] provides an implementation of L_1 and L_2 penalized regression with so-called unpenalized coefficients. In the same vein, the CoxBoost approach [23] forces clinical predictors into the prediction model. In contrast, the PLS+RF approach [25] builds a random forest based on the clinical predictors and "summaries of the molecular predictors" in form of PLS components. Strategy 3 better exploits the predictive potential of clinical predictors than Strategy 1, since they are "favored" in the model building process. In contrast to strategy 2, however, the influence of clinical predictors in the prediction model is affected by molecular predictors. A critical question is how much clinical predictors are/should be favored. Obviously, that should depend on clinical knowledge. It is difficult to give clear recommendations on this heterogeneous family of methods. If clinical predictors are much favored, Strategy 3 is similar to Strategy 2 and the prediction accuracy of the combined model is possibly sub-optimal. If they are not enough favored, however, Strategy 3 has the same pitfall as Strategy 1.

3.4 Strategy 4 ("separate")

Another approach that has sometimes been taken in bioinformatics literature [26] consists in fitting two separate prediction models: one for clinical predictors and one for molecular predictors, and to somehow combine them. This approach, though showing good prediction results in particular situations, has two major pitfalls. First, the clinical predictors are not taken into account when fitting the molecular prediction model: the approach thus fails to focus on residual variation that is not captured by clinical predictors. The second problem is the danger of overfitting through molecular predictors when combining the two models.

3.5 Strategy 5 ("replacement")

Use the molecular data to replace one or more of the "weaker" components of a clinical index. The effect of a component may be "weak" if its relative importance is low [27] or if it is affected by measurement errors. For example, tumor grade is one of three variables in the NPI [3]. As the assessment of tumor grade can depend on the investigator, the general usefulness of NPI may be improved if grade could be replaced by more objective molecular information. It is well-known that variables measured with some subjective component may increase the predictive value in the original data but fail to show its predictive value in new data [28].

4 Validation of the added predictive value

4.1 Why validation?

Validation of prediction models using independent data is important from a clinical point of view, because it measures the accuracy of the prediction model based on a possibly different patient population and thus assesses its generalizability. Model calibration may be required in this context [29]. Good discrimination in new data is an important prerequisite for a good prediction model. In the context of translational research, the *validation of added predictive value* is perhaps even more important than the validation of the prediction accuracy of the prediction model. Some approaches have been proposed for assessing added predictive value based on a single training data while avoiding overfitting problems, see Section 5 for important examples. In this section, however, we address the assessment of added predictive value based on independent validation data. Note that, from a technical point of view, an independent validation data set can be generated artificially from a large data set by random splitting.

The many approaches reviewed below can be classified according to various characteristics. A summary of these important characteristics is given in Table 3. In a nutshell, approach A is based on prediction models, while the other approaches are based on scores only. While approaches A and B consider combined models/scores, the other approaches consider clinical and molecular scores, but no combined scores. In approach C, the assessment of added predictive value is performed through significance testing in multivariate models fitted on the validation data set, while approach D is based on cross-validation or related resampling approaches performed on validation data.

4.2 Validation approaches

4.2.1 Comparing clinical prediction model and combined prediction model on validation data (approach A)

The idea is here to fit two prediction models based on the training data: a clinical prediction model and a combined prediction model. Note that the combined prediction model should be fitted using strategy 2 or strategy 3. Otherwise the results cannot be correctly interpreted. The two models are then applied to make a prediction for the observations from the validation data set, and the predicted and true outcomes are compared for both models. Depending on the type of outcome (right-censored time-to-event or class) and on the point of view of the researcher, different assessment criteria are available.

For class prediction misclassification tables can be computed, and the specificity, sensitivity and error rate of the two prediction models can be compared using standard statistical tests (test of equality of proportions). See Pencina et al [10] for further procedures. A comprehensive description of summary measures is given by Gu and Pepe [30]. For time-to-event outcomes the (integrated) Brier-score and related methods such as prediction error curves [31, 32] are popular measures, but other may also be used depending on the main focus of the study [33, 34]. Measures based on the Brier score are implemented in the R package **pec** [35]. The problem of the choice of a suitable measure to assess the added value is similar for other approaches like approach D discussed below, that is also based on the accuracy of prediction models.

Note that approach A includes as a special case the scenario where the whole clinical prediction model is already given in the literature instead of being estimated from the training data.

4.2.2 Comparing clinical score and combined score on validation data (approach B)

In some cases, prediction models resulting from the training phase cannot be directly applied to the validation data. For instance, this may be the case if the training data set was collected within a case-control design while the validation data set stems from a population study with a (much) smaller percentage of cases. The probabilities output by the prediction model from the training phase do not make sense for the validation data set. It is then more appropriate to try to validate the discriminative ability of the score underlying the prediction model rather than the prediction model itself. Another example where it does not make sense to apply the prediction model and determine its prediction error in the validation set is when the molecular predictors are measured at a different scale in both data sets. For instance, gene expression may have been measured using microarrays in the training set but with PCR in the validation set. In this case, it makes sense to look at the values of the score in the validation data and its association with the outcome rather than at the accuracy of the prediction model. One then needs criteria to assess and compare the scores underlying the prediction models instead of the prediction models themselves. ROC curves including tests of equality of AUC can be considered in the case of class prediction. For survival analysis the association between the two scores and the outcome can be assessed using Cox regression, for instance based on quantile survival curves or other measures of discriminative ability.

An important aspect of both approaches A and B is that no model is fitted on the clinical predictors of the validation data set. Instead, clinical predictors are taken into account in the training phase through the use of a combined prediction model. This will not be the case in the other approaches reviewed in the rest of Section 4.

4.2.3 Testing the molecular score based on validation data in a multivariate model adjusting for clinical predictors (approach C)

Approaches A and B are not always used in practice, probably because combined prediction models and combined scores are tricky and not yet well-established. Moreover, practitioners often prefer to establish their score in form of a molecular score without reference to the clinical predictors. Last but not least, the required clinical predictors are sometimes not available for the training data.

The rest of Section 4.2 is devoted to procedures that do not necessitate the use of combined scores and ignore the *clinical* training data in the analysis of the molecular data. The training phase outputs solely a molecular score whose added predictive value is then determined in the validation data set, thus taking into account the clinical predictors of the validation data. We suppose from now on that the training phase outputs a molecular score like in Eq.(1). This molecular score has been constructed without taking the clinical variables of the training data into account. We assume that it can be computed for all observations from the validation data set. It is in a way considered as a "new predictor".

The most natural way to assess the score's association with the outcome while adjusting for clinical predictors is to fit a prediction model based on the validation data using the score as well as the clinical predictors as independent variables. One can then perform a suitable test to check whether the regression coefficient of the score differs significantly from zero. Since the score does not overfit the validation data set, this approach is unbiased in the sense that it does not systematically over-estimate the added predictive value of the molecular predictors. However, it tells nothing about the predictive value in terms of prediction error. Furthermore, p-values get smaller with increasing sample size – independently of the gained prediction accuracy. As stated by Altman and Royston [11] *"usefulness is determined by how well a model works in practice, not by how many zeros there are in the associated p-values"*. In other words, small p-values may be observed even if the gained prediction accuracy is poor. For a binary outcome Pepe et al [36] illustrate that odds ratios have to be extremely high (e.g. 10 or more) to improve classification. Even a "large" odds ratio, e.g. 3, does not give sufficient strength for a suitable classification tool. They also discuss this issue in the context of the added value of a marker.

Note that approach C includes as a special case the scenario where an aggregated clinical index (such as the IPI score or the Nottingham index) is already given from the literature. Indeed, such an aggregated score is not different from usual clinical predictors from a statistical point of view.

4.2.4 Comparing prediction models with and without molecular score by cross-validation in validation data (approach D)

Approach D is similar to approach C but is not based on the p-value in multivariate regression, thus addressing the important pitfall of approach C. It consists in comparing – via cross-validation or related resampling methods – the prediction accuracy of prediction models with and without molecular score that are fitted on the validation data set. Note that these prediction models can be constructed via logistic or Cox regression or by any other model building approach.

Like in approach C, the molecular score is considered as a new predictor. While approach C assesses this new predictor based on the p-value obtained in a multivariate regression, approach D explicitly evaluates the gain of accuracy yielded by the new predictor by cross-validation. More precisely, the validation data are divided into a number k of cross-validation folds, for instance $k = 10$. In the kth iteration, the kth fold is excluded from the data and two prediction models are fitted to the remaining $k-1$ folds: one model with clinical predictors only and one model with both the score and the clinical predictors. The two models are applied to the kth fold and evaluated based on a suitable criterion like the Brier-score [31, 32] (for survival analysis), the error rate or the AUC (for class prediction). This approach has the major advantage that it can quantify the accuracy gain obtained by incorporating the score. Note that it is also possible to repeat cross-validation to achieve a higher stability, or to use other resampling schemes such as 0.632 or 0.632+ bootstrapping [37].

4.3 Variants of approaches A, B, C, D

4.3.1 Subgroup analysis (SA)

A variant of the approaches A, B, C, D discussed above consists in repeating the validation analyses in several clinical subgroups. The first step – construction of scores/models using training data – is performed exactly as described in Section 4.2, but the evaluation step using validation data is performed separately for the considered subgroups. This approach allows to identify cases where the molecular data may have added predictive value in some subgroup(s) but not in all. A typical example is when the score is highly significant for intermediate subgroups but does not help for extreme subgroups that are already accurately predicted by clinical predictors. This approach is most often applied to pre-defined subgroups (e.g., positive mutational status). A variant would be to consider subgroups defined from a clinical prediction model fitted on the training data.

If there are few important clinical predictors, it may be possible to consider all possible subgroups successively. For instance, in the extreme case where only one binary predictor is important, e.g., the "mutational status", performing the analyses in the negative status group and in the positive status group successively automatically adjusts for the (only) clinical predictor "mutational status". There are no clinical predictors left, and approaches A, B, C, D can be formulated differently. In approaches A and B the comparison is not between clinical and combined models anymore, but rather between the intercept model and the molecular component of the combined model. Approach C is based on significance testing in univariate regression rather than multivariate regression, since there is only one predictor left – the score. Similarly, in each cross-validation iteration approach D compares the score-based univariate model with an intercept model.

If there are several important clinical predictors, however, one may define the subgroups based on one (or two) particular clinical predictor(s) and then proceed exactly as described above with the remaining clinical predictors and the molecular model/score. However, the sample size is often too small for subgroup analyses.

4.3.2 Subgroup analysis *with different models fitted for each subgroup* (SA2)

In the subgroup analysis procedure SA sketched above, the subgroup structure is completely ignored in the training step, i.e. for the construction of the models/scores based on training data. An interesting variant of subgroup analysis, denoted as SA2 in our paper, would be to fit different models/scores to the considered clinical subgroups in the training phase. This approach would take potential interaction effects between clinical and molecular predictors into account. For instance, a specific marker may be predictive of the further disease course in men but not in women. An approach with a global molecular score for all patients would ignore this difference and probably underevaluate the prediction accuracy of the score in men. An important limitation of this approach in practice is that it requires a large sample size so that the size of the subgroups of interest is sufficient to fit molecular scores. Moreover, it evaluates the added predictive value of several scores/models, which may be not appropriate in translational research where simplicity of the score/model is an important aspect. Note that, like variant SA, this variant can be accommodated to all approaches A, B, C and D.

4.3.3 Clinical score fitted on training data (CSFTD)

Approaches C and D consider multivariate models fitted on the validation data set estimating the effects of individual clinical predictors. A variant would be to fit multivariate models based on only two "aggregated predictors", namely the clinical and molecular scores fitted from the training data.

The difference between this variant and the original version of approaches C and D is that the coefficients of the individual clinical predictors are now fitted based on the training data. In contrast, approaches C and D fit the coefficients of clinical predictors with the validation data – while the components of the score with corresponding weights are estimated in the training data. In a sense, the original approaches C and D may slightly disadvantage the molecular score, especially when there are many clinical predictors. In fact, variant CSFTD considers the problem from a different point of view. In approaches C and D, the molecular score is viewed as a potential new predictor which is treated like clinical predictors in the multivariate regression on the validation data: approaches C and D assess the new "molecular predictor". In contrast, the present variant CSFTD rather assesses the score building processes: both scores are thus fitted with the same data. Variant CSFTD may be particularly interesting from a methodological point of view. In clinical applications, however, the assessment of the model building processes is of moderate interest and the original variants of C and D may be more appropriate because they better correspond to the concept of added predictive value by allowing the contribution of the individual clinical predictors to be affected by the molecular score.

4.3.4 Clinical score given from literature (CSG)

An important variant of the CSFTD approach is obtained when the clinical score is already given from the literature instead of being fitted on the training data. This approach is denoted as CSG (standing for **C**linical **S**core **G**iven). In this approach, the clinical data of the training set are *not* used – in contrast to CSFTD.

5 Added predictive value in training data

While the procedures outlined previously are essentially based on *two* data sets – a training data set and a validation data set – this section is devoted to methods using a single data set. Note that, if this single data set is large enough, it can potentially be split randomly in order to apply the methods reviewed above. From now on, we consider that the data set at hand cannot be split, say, because it is too small or to avoid well-known problems caused by data splitting [38, 39]. Some approaches have been proposed to assess the added predictive value of molecular predictors in this case. Roughly, they can be divided into two categories: the global test approaches with adjustment and the resampling-based approaches.

5.1 Global tests with adjustment

Global tests with adjustment are based on linear models with linear predictor

$$\eta = \beta_0 + \beta_1 Z_1 + \cdots + \beta_q Z_q + \beta_1^* X_1 + \cdots + \beta_p^* X_p.$$

In the example of logistic regression, the linear predictor η is linked to the probabilities of the two classes $Y = 0$ and $Y = 1$ through the logistic function. In Cox regression, the linear predictor corresponds to the hazard ratio. The idea of global tests in the context of prediction is to test the null-hypothesis

$$\beta_1^* = \ldots = \beta_p^* = 0$$

i.e. that $X_1, \ldots, X_p$ have no added predictive value in the considered generalized linear model. The two global tests by Goeman and colleagues [40, 41] and Boulesteix and Hothorn [22] differ in the method used to test this hypothesis. Goeman considers a hierarchical model where the regression coefficients have a prior distribution with variance τ^2 and then test the null-hypothesis $\tau^2 = 0$ based on asymptotic results. In contrast, Boulesteix and Hothorn fit regularized regression models using boosting regression with the clinical score as offset. Note that other regularized regression techniques could be used in place of boosting at this stage. They test the null-hypothesis by permutation of the molecular predictors $X_1, \ldots, X_p$ – while the clinical predictors $Z_1, \ldots, Z_q$ (and thus the offset) remain unchanged.

The approaches by Goeman et al [40, 41] and Boulesteix and Hothorn [22] can be applied both to survival analysis and class prediction. They are implemented in the freely available R packages **globaltest** and **globalboosttest**, respectively. It has been shown in simulations that **globalboosttest** performs somewhat better in the important case of few strong molecular predictors, since boosting regression focuses on good predictors while ignoring the other. Another important global testing approach that can be applied to class prediction is the GlobalAncova method by Hummel et al [42] implemented in the R package **GlobalAncova** [43]. It is based on parallel analyses of variance performed for all molecular predictors simultaneously with the class as factor and allows adjustment for clinical predictors.

A shortcoming of such global approaches in the context of added predictive value is that they provide a test but not a comparison of prediction errors. One may thus face situations where the global test identifies added predictive value (i.e. yields small p-values) but the prediction model based on molecular predictors performs poorly. With respect to the connection between global tests and prediction models, note that there is an essential difference between Goeman's global test and the globalboosttest approach by Boulesteix and Hothorn. While Goeman's globaltest does not refer to a specific prediction model, the globalboosttest can be seen as a permutation test of the model fitted by boosting regression based on the non-permuted data set.

5.2 Resampling approaches

The approaches A and B discussed in Sections 4.2.1 and 4.2.2 can be easily applied in cross-validation settings. At each iteration, the excluded fold plays the role of a validation data set while the other folds are used as training data. Combining the results of the cross-validation iterations for testing purposes, however, may be difficult because the iterations are not independent. In the case where two prediction models are applied at each iteration and a p-value is computed to compare their prediction error, van de Wiel et al [44] propose a procedure to combine the obtained p-values while controlling the type I error.

Another resampling-based approach proposed in the literature to assess added predictive value is the so-called pre-validation [45]. The term pre-validation refers to a cross-validation (CV) performed within the available data set S. At each CV iteration j, a molecular score is derived from the data set $S \setminus S_j$, where S_j stands for the jth CV fold, and then computed for the observations from S_j. Since the folds S_j form a partition of the data set S, one thus obtains a score value for each observation. This score value, denoted as "pre-validated score", is not expected to overfit the data set, since at each CV iteration there is no overlap between the "training data" $S \setminus S_j$ and the fold S_j. Finally, a multivariate regression model is fitted using this pre-validated score and the clinical predictors as predictors. The added predictive value is assessed by testing the significance of the regression coefficient of the score. A problem of this procedure is that the conditions for hypothesis testing are not fulfilled: the observations are not independent of each other, since the pre-validated score is derived based on the other observations. A permutation-based improved pre-validation procedure has been proposed recently [46] to address this issue.

Pre-validation is essentially similar to the approach C reviewed above in that it assesses within multivariate regression a score that has been derived on other data. However, pre-validation cannot be seen as a resampling-based extension of approach C. In approach C, the multivariate regression model is based on the validation data set only. In contrast, pre-validation does not fit multivariate regression models based on the test fold S_j at each iteration. It fits only one single multivariate regression model based on the whole data set S – hence the problem with the observations' mutual dependence. To conclude, we point out that pre-validation, like the global tests reviewed at the beginning of this section, does not assess the gain of accuracy but merely provides a test of significance. Like all tests of significance, it can yield a small p-value even if the gain of accuracy is negligible.

6 Other related evaluation procedures

Many other approaches are conceivable in the context of validation of added predictive value. In this section we briefly outline and discuss some of them that *do not* allow to validate added predictive value in the strict sense – while they might be useful as preliminary or additional analyses.

6.1 Testing of the molecular score without adjustment for clinical predictors

This is of course an important preliminary step and such an analysis should be routinely performed. However, it says nothing about added predicted value, except if this test is performed in a clinical subgroup as discussed in Section 4.3. Note that univariate testing of the score is more likely to yield significance if the score was built without taking clinical predictors into account. That is because, in this case, the score is likely to be highly correlated to the good clinical predictors. In contrast, significance of the score might be a sign of potential added predictive value if the molecular score was fitted, say, through penalized regression with the clinical score as an offset.

6.2 Comparing clinical prediction model and molecular prediction model

If the molecular prediction model performs substantially better than the clinical prediction model, the added predictive value is established. However, this will not be the case [4] in most practical cases, and more sophisticated strategies like those reviewed above are necessary to address the question of added value. Moreover, just comparing the clinical and molecular prediction models does not tell us what could be achieved by combining both types of predictors. Thus, such an analysis answers only one part of the question. A special case where they are more useful is when the researcher aims to establish a molecular score that potentially *replaces* a previous score and gives better separation. In this case, the molecular score is expected to yield good accuracy by itself, i.e. to outperform clinical predictors. Note that this is a much stronger request than added predictive value.

6.3 Univariate testing of genes involved in the prediction rule using validation data

If we have a molecular score like the score of Eq.(1), it may be interesting to look whether each of its components GeneA, GeneB and GeneC are univariately significantly associated with the outcome in the validation data. One may also perform a multivariate analysis (Cox regression or logistic regression) based on GeneA, GeneB and GeneC and check whether they are all significant, and whether the sign of the association is the same as in the score obtained from the training phase. Components of the score that are highly significant in the training data, but not in the validation data may indicate a lack of stability or heterogeneity between the two data sets. However, it should be emphasized that significance of all components in the validation data is not a necessary condition for considering the score as "validated". Significance of all components is even quite unlikely if the score is based on a larger number of genes. Conversely, the score given in Eq.(1) may yield a poor gain of accuracy even if the univariate p-values of GeneA, GeneB and GeneC are smaller than 0.05. Note that a change of sign of the coefficients may suggest that a better score could be obtained by removing this component, especially if this change of sign is observed in the multivariate analysis.

6.4 Comparing prediction models obtained from training and validation data

As an outlook, one may also repeat the model building procedure based on the validation data and compare both prediction models. Note that it requires that the same molecular predictors are available for both training and validation data, which is not always the case in practice (for example because another type of array or another technique like PCR was used for the validation data set).

Such a comparison would be interesting because it relates to the stability of the prediction models. Of course, it would be satisfying to find similar models in training and validation data sets. However, the obtained models are more likely to differ substantially because high-dimensional model building is a very instable process [47, 48]. The

top-ranking predictors in high-dimensional data differ strongly even in the case of overlapping subsamples or bootstrap samples [48]. One can thus not reasonably expect to obtain the same model based on two non-overlapping high-dimensional data sets. For several reasons the models may even differ substantially – even if the model from the training phase is validated.

7 Concluding remarks

In this paper, we have reviewed a number of procedures that can be used to validate added predictive value based on validation data as well as methods to assess added predictive value using a single training data set. It is impossible to generally recommend one of these methods over the other, because some methodological issues need further research and the choice strongly depends on the considered particular situation. However, some specific recommendations to avoid common errors while validating added predictive value are given in Table 4 in form of a dos and don'ts list. Concerning microarray-based prediction models in general and their reporting, some guidance is already available in previous articles [49, 50].

As a conclusion, we emphasize the impressing number of mutually connected approaches to validate added predictive value and the lack of guidelines and standards. As others [51], we feel that, in this context, the term "validation" is sometimes used without enough precisions on the considered specific procedure. More research is needed to establish standardized workflows and evaluate the respective merits of the many possible variants outlined in our review.

Acknowledgments

This work was partially supported by the LMU-innovativ Project BioMed-S: Analysis and Modelling of Complex Systems in Biology and Medicine. We thank Monika Jelizarow for her comments.

References

[1] T.R. Golub, D. K. Slonim, P. Tamayo, and et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286: 531–537, 1999.

[2] The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-hodgkin's lymphoma: report of the jury. *New England Journal of Medicine*, 329:987–994, 1993.

[3] M. H. Galea, R. W. Blamey, and C. E. Elston et al. The nottingham prognostic index in primary breast cancer. *Breast Cancer Research and Treatment*, 22:207–219, 1992.

[4] P. Eden, C. Ritz, and C. Rose et al. "Good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *European Journal of Cancer*, 40:1837–1841, 2004.

[5] C. Truntzer, D. Maucort-Boulch, and P. Roy. Comparative optimism in models involving both classical clinical and gene expression information. *BMC Bioinformatics*, 9:434, 2008.

[6] R. Simon. Development and validation of therapeutically relevant multi-gene biomarker classifiers. *Journal of the National Cancer Institute*, 97:866–867, 2006.

[7] M. Buyse, S. Loi, and L. van't Veer et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *Journal of the National Cancer Institute*, 98:1183–1192, 2006.

[8] S. George. Statistical issues in translational cancer research. *Clinical Cancer Research*, 14:5954–5958, 2008.

[9] J. P. A. Ioannidis. Expectations, validity, and reality in omics. *Journal of Clinical Epidemiology*, 63:960–963, 2010.

[10] M. J. Pencina, R. B. D'Agostino, R. B. D'Agostino, and R. S. Vasan. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27:157–172, 2008.

[11] D. Altman and P. Royston. What do we mean by validating a prognostic model? *Statistics in Medicine*, 19:453–473, 2000.

[12] P. Royston and W. Sauerbrei. *Multivariable model-building*. Wiley, New York, 2008.

[13] A. L. Boulesteix and C. Strobl. Optimal classifier selection and negative bias in error rate estimation: An empirical study on high-dimensional prediction. *BMC Medical Research Methodology*, 9:85, 2009.

[14] S. E. Bleeker, H. A. Moll, and E. W. Steyerberg et al. External validation is necessary in prediction research: a clinical example. *Journal of Clinical Epidemiology*, 56: 826–832, 2003.

[15] I. R. König, J. D. Malley, and C. Weimar et al. Practical experiences on the necessity of external validation. *Statistics in Medicine*, 26:5499–5511, 2007.

[16] R. Tibshirani. Regression shrinkage and selection via the LASSO . *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

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

[18] A. Benner, M. Zucknick, T. Hielscher C. Ittrich, and U. Mansmann. High-dimensional cox models: the choice of penalty as part of the model building process. *Biometrical Journal*, 52:50–69, 2010.

[19] E. Bair and R. Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, 2:0511, 2004.

[20] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

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

[22] A. L. Boulesteix and T. Hothorn. Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics*, 11:78, 2010.

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

[24] J. J. Goeman. *penalized: L1 (lasso) and L2 (ridge) penalized estimation in GLMs and in the Cox model*, 2010. R Package version 0.9-31.

[25] A.-L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer. Evaluating microarray-based classifiers: an overview. *Cancer Informatics*, 6:77–97, 2008.

[26] O. Gevaert, F. de Smet, and D. Timmermann et al. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, 22:e184–e190, 2006.

[27] M. Schemper. The relative importance of prognostic factors in studies of survival. *Statistics in Medicine*, 12:2377–2382, 1993.

[28] T. L. Diepgen, W. Sauerbrei, and M. Fartasch. Development and validation of diagnostic scores for atopic dermatitis incorporating criteria of data quality and practical usefulness. *Journal of Clinical Epidemiology*, 49:1031–1038, 1996.

[29] H. van Houwelingen. Validation, calibration, revision and combination of prognostic survival models. *Statistics in Medicine*, 19:3401–3415, 2000.

[30] Gu and M. Pepe. Measures to summarize and compare the predictive capacity of markers. *International Journal of Biostatistics*, 5:27, 2009.

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

[32] T. Gerds and M. Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48: 698–705, 2006.

[33] P. Royston and W. Sauerbrei. A new measure of prognostic separation in survival data. *Statistics in Medicine*, 23:723–748, 2004.

[34] E. W. Steyerberg, A. J. Vickers, and N. R. Cook et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21:128–138, 2010.

[35] T. A. Gerds. *pec: Validation of predicted surival probabilities using inverse weighting and resampling*, 2009. R Package version 1.1.1.

[36] M. S. Pepe, H. Janes, and G. Longton et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159:882–890, 2004.

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

[38] R. P. Hirsch. Validation samples. *Biometrics*, 47:1193–1194, 1991.

[39] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95:14–18, 2003.

[40] J. Goeman, S. van de Geer, F. de Kort, and H. C. van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20: 93–99, 2004.

[41] J. J. Goeman, J. Oosting, and A. M. Cleton-Jansen et al. Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21:1950–1957, 2005.

[42] M. Hummel, R. Meister, and U. Mansmann. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, 24:78–85, 2008.

[43] U. Mansmann, R. Meister, M. Hummel, and R. Scheufele. *GlobalAncova: Calculates a global test for differential gene expression between groups*, 2010. R Bioconductor Package version 3.16.0.

[44] M. van de Wiel, J. Berkhof, and W. van Wieringen. Testing the prediction error difference between 2 predictors. *Biostatistics*, 550–560:10, 2009.

[45] R. Tibshirani and B. Efron. Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, 1:1, 2002.

[46] H. Höfling and R. Tibshirani. A study of pre-validation. *Annals of Applied Statistics*, 2:643–664, 2008.

[47] L. Ein-Dor, I. Kela, and G. Getz et al. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21:171–178, 2005.

[48] A. L. Boulesteix and M. Slawski. Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10:556–568, 2009.

[49] E. E. Ntzani and J. P. A. Ioannidis. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *The Lancet*, 362:1439–1444, 2003.

[50] A. Dupuy and R. Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute*, 99:147–157, 2007.

[51] J. M. Taylor, D. P. Ankerst, and R. R. Andridge. Validation of biomarker-based risk prediction models. *Clinical Cancer Research*, 14:5977–5983, 2008.

Table 1: Glossary with examples

	Definition	Example
score	risk index derived from the training set	$-0.14 \times$ Sex $- 0.11 \times$ geneA $+ 0.21 \times$ geneB $+ 0.09 \times$ geneC
clinical score	score involving clinical predictors only	$-0.14 \times$ Sex $+ 0.02 \times$ Age
molecular score	score involving molecular predictors only	$- 0.11 \times$ geneA $+ 0.21 \times$ geneB
prediction model (PM)	function assigning a new observation to a class	$\hat{Y} = 1$ if score> 0.2 $\hat{Y} = 0$ otherwise
clinical PM	A PM based on clinical predictors only	
molecular PM	A PM based on molecular predictors only	

This table gives synthetic definitions of important terms.

Table 2: Combined prediction models – Overview

	1: naive	**2: residuals**	**3: favoring**	**4: separate**	**5: replacement**
one-step approach	yes	no	yes	no	no
treats clinical and molecular predictors equally	yes	no	no	no	no
essentially depends on a crucial parameter	no	no	yes	no	no
contribution of clinical predictors is affected by molecular predictors	yes	no	yes	yes	depends
fits an only-clinical model	no	yes	no	yes	yes
fits a molecular model to the residuals of the clinical model from 1st step	no	yes	no	no	no
aims to replace a problematic clinical component through molecular data	no	no	no	no	yes
adequate to assess added predictive value	no	yes	depends	depends	depends

This table gives a summary of the five strategies reviewed in the introduction to build combined prediction models based on both clinical and molecular predictors.

Table 3: Assessing added predictive value – Overview

	A	B	C	D				
uses combined models/scores	yes	yes	no	no				
is based on scores only	no	yes	yes	yes				
is based on the accuracy gain as estimated directly on validation data	yes	yes	no	no				
is based on the accuracy gain as estimated through resampling on validation data	no	no	no	yes				
is based on significance testing in multivariate models fitted on validation data	no	no	yes	no				
considers the molecular score as a "new predictor"	no	no	yes	yes				
fits (a) model(s) to the validation data	no	no	yes	yes				
fits a model to clinical data of the training set	yes	yes	no	no				
					SA variant	SA2 variant	CSFTD	GSG
performs subgroup analyses					yes	yes		
accounts for interactions between clinical and molecular predictors					no	yes		
fits the clinical score based on training data							yes	no

This table gives a summary of the important characteristics of procedures A,B,C,D and variants SA, SA2, CSFTD and CSG reviewed in Section 5.

Table 4: DOs and DONTs

DON'T	modify the score or the prediction model after seeing the results on the validation data set.
DON'T	select the cutpoint to dichotomize the score based on validation data.
DO	select a *unique* score/prediction model for each setting (only clinical, only molecular or combined – depending on the adopted approach) *before* opening the validation data set.
DO	also assess the added predictive value based on other criteria than p-values, because p-values may be small even if the accuracy gain is not relevant from a biomedical point of view.
DON'T	fit a combined model by considering all predictors equally: the assessment of added predictive value of molecular data is essentially asymmetric.
DON'T	only consider error rates in the case of class prediction. ROC or related approaches are also useful.
DON'T	think that statistically valid tests for assessing accuracy gain can be derived by considering cross-validation or bootstrap iterations as statistical units. The iterations are not independent and this has to be taken into account.
DO	keep in mind that a p-value from a misspecified model cannot be interpreted. In contrast, the comparison of accuracy through a correct validation scheme (e.g. cross-validation) is interpretable even if the underlying prediction models are misspecified.

This table gives a list of general recommendations on the assessment of added predictive value.