BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 3. 215–229
Advance Access published on 18 January 2011
doi:10.1093/bib/bbq085

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

Anne-Laure Boulesteix and Willi Sauerbrei

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 or an established index 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.

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

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 10 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, that are of variable utility and often not 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. Clinical predictors may be given as a list of individual factors or in form of a well-established index such as the International Prognostic Index (IPI) for lymphoma [2] or the Nottingham Prognostic Index (NPI) for breast cancer [3]. In this article, 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.

Note that, in particular cases, researchers are interested in molecular predictors that would be able to

Corresponding author. Anne-Laure Boulesteix, Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany. Tel: +49-89-7095-7598; Fax: +49-89-7095-7491; E-mail: boulesteix@ibe.med.uni-muenchen.de

Anne-Laure Boulesteix, PhD, is an assistant professor of computational molecular medicine at the University of Munich (Germany), where she obtained her PhD in statistics in 2005. Statistical methods in bioinformatics are her main research interest, including prediction models with high-dimensional biological data and validation in bioinformatics.

Willi Sauerbrei, PhD, is a senior statistician and professor in medical biometry at the University Medical Center Freiburg (Germany). He has been working for more than two decades as an academic biostatistician and has authored many research papers in leading statistical and clinical journals. Together with Patrick Royston (London), he has written a book on multivariable model building.

© The Author 2011. Published by Oxford University Press. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

replace classical clinical predictors. For instance, existing markers might also be very expensive to measure, necessitate harmful interventions or involve a subjective component like pathological judgments. In all these cases, the molecular predictors may replace the clinical predictors even if the predictive value is not better. The concept of added predictive value reviewed in this article is then irrelevant, because the molecular predictors actually add something else that cannot be measured on the same scale. Note that, in this case, statistical validation is easier, since it does not necessitate to adjust for existing predictors. From now on, we focus on the frequent situations where classical clinical predictors are reasonably satisfying and need to be completed rather than replaced by molecular predictors.

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. The danger of overfitting when using high-dimensional data is now well-acknowledged in the literature. However, as far as their added predictive power compared with clinical predictors is concerned, overoptimistic conclusions [4, 5] are still a major issue in current research. 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 sufficient. It often yields a dramatic overestimation 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 [6–10] 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, the assessment of the discriminative ability of a derived score 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 molecular model is not sufficient as the clinical interest centers around the added value compared with previous existing models [11]. 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 article discusses statistical techniques and gives some recommendations for practical applications when clinical data and high-throughput data are available.

Note that Altman and Royston [12] 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 [13] but beyond the scope of this article. The focus of this review is on statistical validity.

From now on, we assume 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 Prediction models section 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 [14, 15].

In practice, validation data sets are often data collected later (temporal validation) or data collected elsewhere (external validation) [12]. But the training and test sets may also be drawn randomly from a single data set at hand. Validation using external data (e.g. from a different hospital) is generally considered as stronger than validation based on a randomly selected subset from the data set. Already from traditional research the necessity of external validation is well known [16, 17]. It is the only way to ensure that the derived model may be widely useful and does not only work in the particular setting in which it was developed. Re-calibration may be required [18].

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

If no external data are available, validation based on a randomly selected subset is recommended, because keeping a validation data set unopened until the validation phase is the only way to warrant that the analyst does not consciously or subconsciously take into account the validation data set to, e.g. choose some model parameters, choose a particular variant of a method that is found to work better, etc. In cross-validation (CV) settings, validation unfortunately tends to be 'incomplete' in practice in the sense that not all steps of model selection and choice are cross-validated [19, 20]. For instance, the researcher might CV several classifiers and finally choose the classifier yielding the smallest cross-validation error rate. In this context, Dupuy and Simon [20] recommend to 'report the [cross-validation] estimates for all the classification algorithms if several have been tested, not just the most accurate' to avoid the substantial optimization bias quantitatively assessed by Boulesteix and Strobl [14].

In this sense, a unique splitting into training and validation data set may be advantageous provided the data set is large enough. Note that in survival data sets, the 'sample size' to consider is actually the number of events rather than the number of patients. Although most current high-dimensional molecular data sets are actually too small to be split, the benefits of validation on independent data (especially the substantial reduction of optimization bias) often make up for the inconveniences (smaller training set, dependence on the particular split) in many cases. Note that in this case the random split should be performed at the very beginning of the statistical analysis. The researchers should not repeat the analysis with several splits successively in a trial-and-error strategy to select a particularly favorable split.

The rest of the article is structured as follows. Prediction models section introduces the terminology of prediction models used in this article. The section 'Strategies to derive combined prediction models' section gives an overview of possible methods for deriving combined prediction models based on both clinical and molecular predictors. Validation of the added predictive value section presents several existing approaches to validate added predictive value based on training and validation data sets, while Added predictive value in training data section briefly reviews procedures like global tests that can be applied—but no limited—to the case of a unique data set. Before the concluding remarks, we will discuss further evaluation procedures in Other related evaluation procedures section.

PREDICTION MODELS

In this article, 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 earlier situation, 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 substantially. Classical clinical predictors such as age, sex, or tumor stage 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 (often a metric scale), the clinical variables $Z_1, \ldots, Z_q$ may be categorical (e.g. sex, tumor stage, estrogens receptor 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 article, 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 an important example in practice. For instance, a 3-genes linear score may be given as

$$\begin{aligned}\text{score} = &-0.113 \times \text{geneA} + 0.207 \times \text{geneB} \\ &+ 0.091 \times \text{geneC},\end{aligned} \tag{1}$$

where 'geneA', 'geneB' and 'geneC' stand for the respective expression levels of these genes.

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

Table 1: Glossary with examples

	Definition	Example
Score	Risk index derived from the training set	$-0.14 \times \text{Sex} - 0.11 \times \text{geneA} + 0.21 \times \text{geneB} + 0.09 \times \text{geneC}$
Clinical score	Score involving clinical predictors only	$-0.14 \times \text{Sex} + 0.02 \times \text{Age}$
Molecular score	Score involving molecular predictors only	$-0.11 \times \text{geneA} + 0.21 \times \text{geneB}$
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	$\log(P(Y=1)/P(Y=0)) = -0.14 \times \text{Sex} + 0.02 \times \text{Age}$
Molecular PM	A PM based on molecular predictors only	$\log(P(Y=1)/P(Y=0)) = -0.11 \times \text{geneA} + 0.21 \times \text{geneB}$

This table gives synthetic definitions of important terms including toy examples in the context of binary class prediction.

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 [21], elastic nets [22, 23], SuperPC [24] or Cox regression performed after univariate filtering. Clinical scores are usually derived with standard variable selection methods using logistic regression or Cox regression. Widely speaking, estimated class probabilities returned by ensemble methods like random forests [25] 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 Equation (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.

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.

Strategy 1 ('naive')

The perhaps most naive 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 (see [26] for an example). 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).

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 [21] or boosting regression [27]. This approach yields a linear predictor in which the coefficients of the clinical variables are prone to selection bias if variable selection is performed [13]

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

Table 2: Combined prediction models—overview

	1: naive	2: residuals	3: favoring	4: dimension reduction	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	Yes	No
Contribution of clinical predictors is affected by molecular predictors	Yes	No	Yes	Yes	Depends
Fits an only-clinical model	No	Yes	No	No	Yes
Fits a molecular model to the residuals of the clinical model from 1st step	No	Yes	No	No	No
Replaces 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.

but are not affected by the molecular predictors. It is adequate to test added predictive value [28] since the focus is here on the residual variation of the outcome. However, it may be suboptimal 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 [29].

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.

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. A comparative study of some of these approaches is given in Bovelstad *et al.* [30] in the context of survival prediction.

For instance, clinical predictors might be favored in terms of prior in Bayesian settings or through a different penalty in penalized regression. The R package penalized [31] provides an implementation of L_1 and L_2 penalized regression with the so-called unpenalized coefficients. Such methods, in particular L_2 penalized regression, have been shown to perform well in terms of prediction in a comparative study on survival prediction from combined models [30]. In the same vein, the CoxBoost approach [29] forces clinical predictors into the prediction model.

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 the 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 suboptimal. If they are not enough favored, however, Strategy 3 has the same pitfall as Strategy 1.

Strategy 4 ('dimension reduction')

Dimension reduction approaches constitute an important special case of methods favoring clinical predictors. We are considering them separately here.

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

They include methods like the PLS + RF procedure (standing for Partial Least Squares followed by Random Forest) or the supervised principal component approach [24] and are based on two successive steps. The molecular predictors are first summarized in form of new components in a dimension reduction step. A prediction rule including these new components and the clinical predictors as covariates is then built using, e.g. the classical Cox model [30] or random forests [32]. A critical aspect of these methods is overfitting that should be avoided while constructing the new components. For instance, over-fitting can be avoided through a pre-validation procedure applied to the dimension reduction step [32, 33]. Otherwise, the new components are likely to be strongly correlated with the outcome even in the case of non-informative molecular predictors, and thus yield suboptimal combined models.

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 [34] or if it is affected by measurement errors. For example, tumor grade is one of three variables in the Nottingham Prognostic Index (NPI) [3]. As the assessment of tumor grade can depend on the investigator, the general usefulness of NPI may be improved if the grade could be replaced by more objective molecular information. It is well known that the 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 [35].

VALIDATION OF THE ADDED PREDICTIVE VALUE

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 [18]. Good discrimination in new data is an important pre-requisite 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 Added predictive value in training data section 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. Compared to data from a new setting (external data), this internal validation approach has disadvantages. External validation is a more stringent procedure necessary for evaluating whether the predictive model will generalize to populations other than the one on which it was developed [36].

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, approaches A and D are based on prediction models, while the other approaches are based on scores only and usually consider the discriminative ability. 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 CV or related resampling approaches performed on validation data.

Validation approaches

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, 3 or 4. 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 time-to-event outcomes, the (integrated) Brier score and related methods such as prediction error curves [37, 38] are popular measures, but others

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

Table 3: Assessing added predictive value—overview

Approaches	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				
Variants					**SA**	**SA2**	**CST**	**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 approaches A,B,C,D and approach variants SA, SA2, CST and CSG reviewed in Added predictive value in training data section.

may also be used depending on the main focus of the study [39, 40]. Measures based on the Brier score are implemented in the R package pec [41]. 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. 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. The Brier score [37] that is often used in survival analysis may also be useful in the context of class prediction. A comprehensive description of summary measures is given by Gu and Pepe [42].

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.

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. In this case, re-calibration may be considered [18]. Another option is to validate the discriminative ability of the score underlying the prediction model rather than the prediction model itself.

A perhaps more characteristic example for which it makes sense to consider the discriminative ability instead of the prediction is the case of molecular predictors that are measured at a different scale in the training and validation data sets. For instance, gene expression may have been measured using microarrays in the training set but using the low-throughput reverse transcription quantitative polymerase chain reaction (RT–PCR) technique in the validation set. The unit of measurement is then not the same for the two data sets. It thus makes no sense to apply the model coefficients derived from the training set to the validation set.

In this case, it may be useful 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. Receiver operating characteristic (ROC) curves including tests of equality of the area under the curve (AUC) or the c-index can be considered in

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

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. As we will discuss later in more details, approaches based directly on prediction accuracy (like approach A) are generally to be preferred to approaches based solely on discriminative ability (like approach B). In this sense, approach B should be seen as a suboptimal variant of approach A to be used in the cases where approach A cannot be applied even after re-calibration.

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 the Validation of the addded predictive value section.

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

Approaches A and B are not widely 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 the form of a molecular score that does not involve clinical predictors. Last but not least, the required clinical predictors are sometimes not available for the training data. The rest of Validation approaches section is devoted to procedures that do not necessitate the use of combined scores. The training phase outputs solely a molecular score like in Equation (1) whose added predictive value is then determined in the validation data set, thus taking into account the clinical predictors of the validation data.

This molecular score may have been constructed while taking the clinical predictors of the training data into account or not. The SuperPC approach [24] is an example of method deriving a molecular score while taking the clinical predictors into account. The idea is to derive the molecular score by applying principal component analysis to predictors that are correlated with the outcome in the training data after adjustment for clinical predictors [30].

No matter how the molecular score is derived, 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 molecular 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 overestimate the added predictive value of the molecular predictors. It has been widely used in prognostic studies involving high-dimensional molecular data [43, 44].

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 [12] '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.* [45] illustrate that in the case of binary classification the odds ratio of a binary marker has to be extremely high (e.g. 10 or more) in order to improve the performance of a classification rule substantially. 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 or the NPI) is already given from the literature. Indeed, such an aggregated score is not different from the usual clinical predictors from a statistical point of view.

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

Approach D is similar to approach C, but it consists in comparing the prediction accuracy of prediction models with and without molecular score via CV or related resampling methods rather than via significance testing. Therefore, it addresses the important pitfall of approach C that was based on *P*-values only. Note that these prediction models can be constructed via logistic or Cox regression or by any other model building approach.

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

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 CV. More precisely, the validation data are divided into a number k of CV folds, for instance $k = 10$. In the k-th iteration, the k-th 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 k-th fold and evaluated based on a suitable criterion like the Brier score [37, 38] (for both survival analysis and class prediction), error rate or AUC (for class prediction only). 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 CV to achieve a higher stability, or to use other resampling schemes such as '0.632' or '0.632+' bootstrapping [46].

Variants of approaches A, B, C, D

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 Validation approaches section, 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. 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 simplify in an obvious way.

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 the subgroup analyses.

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

In the subgroup analysis (SA) 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 SA, denoted as SA2 in this article, 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 inappropriate 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.

Clinical score fitted on training data

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 clinical score

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

fitted on training data (CST) 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 CST rather assesses the score building processes. Both scores are thus fitted with the same data. Variant CST 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.

Clinical score given from literature

A variant of the CST approach with important practical relevance is obtained when the clinical score is taken from the literature instead of being fitted on the training data. This approach is denoted as CSG (standing for Clinical Score Given). In this approach, the clinical data of the training set are *not* required—in contrast to CST.

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 [19, 47]. 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.

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

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

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 [48, 49] and Boulesteix and Hothorn [28] differ in the method used to test this hypothesis. Goeman *et al.* consider 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.* [48, 49] and Boulesteix and Hothorn [28] 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.* [50] implemented in the R package GlobalAncova [51]. 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. Global tests can be seen as an insufficient, but conceptually simple, easily applicable and statistically sound approach to

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

'get an idea' of the markers' potential added predictive value. Note that their use is not restricted to the case of a unique data set. They can be applied to any data set including a validation data set.

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. The globalboosttest can be seen as a permutation test of the model fitted by boosting regression based on the non-permuted data set.

Resampling approaches

The approaches A and B discussed in 'Comparing clinical prediction model and combined prediction model on validation data' (approach A) and 'Comparing clinical score and combined score on validation data' (approach B) sections can be easily applied in CV settings, e.g. *k*-fold CV. At each iteration, the excluded fold plays the role of a validation data set while the other folds are used as training data. Alternatively, one can also consider multiple splits into training and validation data sets. This is very similar to CV, except that the splits do not result from a unique partition of the data set.

Note that such resampling approaches are particularly relevant when the unique data set at hand is not large enough to be split into a training and validation data set, the usual situation with high-dimensional data. If the data set is large enough, a single splitting may be preferred. Keeping the validation data set unopened during the training phase is indeed the only way to warrant that the analyst does not (consciously or subconsciously) use information from the validation data, for instance to select the method's parameters or variant. Keeping the validation data set unopened is of course impossible if several splits into training and validation data set are considered successively. In a word, resampling approaches, if used correctly, can advantageously attenuate the high variability, which characterizes model selection and evaluation in very small sample settings. However, naïve applications can result in biased estimates [14, 20, 52].

Combining the results of the CV/splitting 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.* [53] 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 [33]. The term pre-validation refers to a 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 j-th 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 because 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 [54] 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 and 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.

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 validation of added predictive value in the strict sense—while they might be useful as preliminary or additional analyses.

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

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 Variants of approaches A, B, C, D section. 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, univariate 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.

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.

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

If we have a molecular score like in Equation (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 Equation (1) may yield a poor gain of accuracy even if the univariate *P*-values of GeneA, GeneB and GeneC are smaller than 0.01 [55]. 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.

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 [56, 57]. The top-ranking predictors in high-dimensional data differ strongly even in the case of overlapping subsamples or bootstrap samples [57]. 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.

CONCLUDING REMARKS

In this article, 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

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

Table 4: DOs and DON'Ts

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 CV 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. CV) 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.

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 [20, 58, 36] and some of the other references cited.

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 [36], 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.

Key Points

- Validation of added predictive value of molecular data is more complicated than simple validation of a molecular prediction model.
- A useful option is to build both a clinical model/score and combined model/score based on training data and compare their performance on validation data. There are different possibilities to build a combined model in this context.
- Another recommended option is to assess the accuracy gain yielded by the molecular score compared with the clinical model through CV within the validation data set.
- Methods based solely on *P*-values are not sufficient to assess added predictive value, because small *P*-values can be obtained even if the gain of accuracy is negligible.
- Evaluation of predictive models and added value by using external data is more important as it generalizes to populations other than the one on which the model was developed.

Acknowledgement
We thank Monika Jelizarow for her comments.

FUNDING
This work was partially supported by the LMU-innovativ Project BioMed-S: Analysis and Modelling of Complex Systems in Biology and Medicine.

References

1. Golub TR, Slonim DK, Tamayo P, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;**286**:531–7.
2. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive non-Hodgkin's lymphoma: report of the jury. *N Engl J Med* 1993;**329**:987–94.
3. Galea MH, Blamey RW, Elston CE, *et al*. The Nottingham prognostic index in primary breast cancer. *Breast Cancer Res Treat* 1992;**22**:207–19.
4. Eden P, Ritz C, Rose C, *et al*. "Good old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Cancer* 2004;**40**: 1837–41.
5. Truntzer C, Maucort-Boulch D, Roy P. Comparative optimism in models involving both classical clinical and gene expression information. *BMC Bioinformatics* 2008;**9**: 434.
6. Simon R. Development and validation of therapeutically relevant multi-gene biomarker classifiers. *J Natl Cancer Inst* 2006;**97**:866–7.
7. Buyse M, Loi S, van't Veer L, *et al*. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006;**98**: 1183–92.
8. George S. Statistical issues in translational cancer research. *Clin Cancer Res* 2008;**14**:5954–8.

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

9. Ioannidis JPA. Expectations, validity, and reality in omics. *J Clin Epidemiol* 2010;**63**:960–3.
10. Mischak H, Allmaier G, Apweiler R, *et al*. Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Trans Med* 2010;**2**:1–5.
11. Pencina MJ, D'Agostino RB, D'Agostino RB, *et al*. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med* 2008;**27**:157–72.
12. Altman D, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;**19**:453–73.
13. Royston P, Sauerbrei W. *Multivariable model-building – A pragmatic approach to regression analysis baded on fractional polynomials for modelling continuous variables*. Chichester: Wiley & Son Ltd, 2008.
14. Boulesteix AL, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. *BMC Med Res Methodol* 2009;**9**:85.
15. Jelizarow M, Guillemot V, Tenenhaus A, *et al*. Over-optimism in bioinformatics: an illustration. *Bioinformatics* 2010;**26**:1990–8.
16. Bleeker SE, Moll HA, Steyerberg EW, *et al*. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003;**56**:826–32.
17. König IR, Malley JD, Weimar C, *et al*. Practical experiences on the necessity of external validation. *Stat Med* 2007;**26**: 5499–511.
18. van Houwelingen H. Validation, calibration, revision and combination of prognostic survival models. *Stat Med* 2000; **19**:3401–15.
19. Simon R, Radmacher MD, Dobbin K, *et al*. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 2003;**95**:14–8.
20. Dupuy A, Simon R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* 2007;**99**: 147–57.
21. Tibshirani R. Regression shrinkage and selection via the LASSO. *J Royal Stat Soc B* 1996;**58**:267–88.
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc B* 2005;**67**:301–20.
23. Benner A, Zucknick M, Hielscher T, *et al*. High-dimensional cox models: the choice of penalty as part of the model building process. *Biomet J* 2010;**52**:50–69.
24. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol* 2004; **2**:0511.
25. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
26. Hothorn T, Bühlmann P, Dudoit S, *et al*. Survival ensembles. *Biostatistics* 2006;**7**:355–73.
27. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting (with discussion). *Stat Sci* 2007;**22**:477–505.
28. Boulesteix AL, Hothorn T. Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinformatics* 2010;**11**:78.
29. Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics* 2008;**9**:14.
30. Bovelstad HM, Nygard S, Borgan O. Survival prediction from clinico-genomic models – a comparative study. *BMC Bioinformatics* 2009;**10**:413.
31. Goeman JJ. *Penalized: L1 (Lasso) and L2 (Ridge) Penalized Estimation in GLMs and in the Cox Model,* 2010. R Package version 0.9-31.
32. Boulesteix A-L, Strobl C, Augustin T, *et al*. Evaluating microarray-based classifiers: an overview. *Cancer Informat* 2008;**6**:77–97.
33. Tibshirani R, Efron B. Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol* 2002;**1**:1.
34. Schemper M. The relative importance of prognostic factors in studies of survival. *Stat Med* 1993;**12**:2377–82.
35. Diepgen TL, Sauerbrei W, Fartasch M. Development and validation of diagnostic scores for atopic dermatitis incorporating criteria of data quality and practical usefulness. *J Clin Epidemiol* 1996;**49**:1031–8.
36. Taylor JM, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. *Clin Cancer Res* 2008;**14**:5977–83.
37. Graf E, Schmoor C, Sauerbrei W, *et al*. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med* 1999;**18**:2529–45.
38. Gerds T, Schumacher M. Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biomet J* 2006;**48**:698–705.
39. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004;**23**:723–48.
40. Steyerberg EW, Vickers AJ, Cook NR, *et al*. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;**21**: 128–38.
41. Gerds TA. *pec: Validation of Predicted Surival Probabilities Using Inverse Weighting and Resampling.* 2009. R Package version 1.1.1.
42. Gu W, Pepe MS. Measures to summarize and compare the predictive capacity of markers. *Int J Biostat* 2009;**5**:27.
43. Metzeler KH, Hummel M, Bloomfield CD, *et al*. An 86 probe set gene expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* 2008; **112**:4193–201.
44. Yao M, Huang Y, Shioi K, *et al*. A three-gene expression signature model to predict clinical outcome of clear cell renal carcinoma. *Int J Cancer* 2008;**123**:1126–32.
45. Pepe MS, Janes H, Longton G, *et al*. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;**159**:882–90.
46. Schumacher M, Binder H, Gerds T. Assessment of survival prediction models based on microarray data. *Bioinformatics* 2007;**23**:1768–74.
47. Hirsch RP. Validation samples. *Biometrics* 1991;**47**:1193–94.
48. Goeman JJ, van de Geer SA, de Kort F, *et al*. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004;**20**:93–9.
49. Goeman JJ, Oosting J, Cleton-Jansen AM, *et al*. Testing association of a pathway with survival using gene expression data. *Bioinformatics* 2005;**21**:1950–7.
50. Hummel M, Meister R, Mansmann U. GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics* 2008;**24**:78–85.

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019

51. Mansmann U, Meister R, Hummel M, *et al*. *GlobalAncova: Calculates a Global test for Differential Gene Expression Between Groups.* 2010. R Bioconductor Package version 3.16.0.

52. Lusa L, McShane LM, Rademacher MD, *et al*. Appropriateness of some resampling-based inference procedures for assessing performance of prognostic classifiers derived from microarray data. *Stat Med* 2007;**26**:1102–13.

53. van de Wiel M, Berkhof J, van Wieringen W. Testing the prediction error difference between 2 predictors. *Biostatistics* 2009;**10**:550–60.

54. Höfling H, Tibshirani R. A study of pre-validation. *Ann App Stat* 2008;**2**:643–64.

55. Pepe MS, Janes H, Longton G, *et al*. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;**159**:882–90.

56. Ein-Dor L, Kela I, Getz G, *et al*. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* 2005;**21**:171–8.

57. Boulesteix AL, Slawski M. Stability and aggregation of ranked gene lists. *Brief Bioinformatics* 2009;**10**:556–68.

58. Ntzani EE, Ioannidis JPA. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 2003;**362**:1439–44.

Downloaded from https://academic.oup.com/bib/article-abstract/12/3/215/258290 by guest on 30 November 2019