On the Simultaneous Analysis of Clinical and Omics Data - a Comparison of Globalboosttest and Pre-validation Techniques

Margret-Ruth Oelker and Anne-Laure Boulesteix

Abstract In medical research biostatisticians are often confronted with supervised learning problems involving different kinds of predictors including, e.g., classical clinical predictors and high-dimensional "omics" data. The question of the *added* predictive value of high-dimensional omics data given that classical predictors are already available has long been under-considered in the biostatistics and bioinformatics literature. This issue is characterized by a lack of guidelines and a huge amount of conceivable approaches. Two existing methods addressing this important issue are systematically compared in the present paper. The *globalboosttest* procedure (Boulesteix and Hothorn, 2010) examines the additional predictive value of high-dimensional molecular data via boosting regression including a clinical offset, while the *pre-validation* method sums up omics data in form of a new cross-validated predictor that is finally assessed in a standard generalized linear model (Tibshirani and Efron, 2002). Globalboosttest and pre-validation are introduced and discussed, then assessed based on a simulation study with survival data and finally applied to breast cancer microarray data for illustration. R codes to reproduce our results and figures are available from http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/gbtpv/index.html

Key words: globalboosttest, pre-validation, cross-validation, added predictive value, survival analysis, boosting, permutation tests

Margret-Ruth Oelker
Department of Statistics, University of Munich, Ludwigstr. 33, 80539 Munich, Germany / Department of Medical Informatics, Biometry and Epidemiology of the Faculty of Medicine, University of Munich, Marchioninistr. 15, 81377 Munich, Germany e-mail: margret.oelker@stat.uni-muenchen.de

Anne-Laure Boulesteix
Department of Medical Informatics, Biometry and Epidemiology of the Faculty of Medicine, University of Munich, Marchioninistr. 15, 81377 Munich, Germany e-mail: boulesteix@ibe.med.uni-muenchen.de

1 Introduction

While high-dimensional "omics" data such as microarray transcriptomic data have been studied in the context of outcome prediction for more than ten years in biomedical research, the question of the added predictive value of such data given that classical predictors are already available has comparatively focused less attention in the literature (e.g. Boulesteix and Sauerbrei, 2011). For a given prediction problem (for example prediction of response to therapy or survival time), we often have two types of predictors. On the one hand, conventional clinical covariates such as, e.g. age, sex, disease duration or tumour stage are available as potential predictors. They have been typically extensively investigated and validated in previous studies. On the other hand, we have a large number of "omics" predictors that are generally more difficult to measure than classical clinical predictors and not yet well-established. In the context of translational biomedical research, researchers are interested in the added predictive value of such omics predictors over classical clinical predictors.

The combined analysis of high-dimensional omics predictors and low-dimensional clinical predictors raises various challenges. How can we build a combined model that is optimal in terms of prediction accuracy? How can we test the added predictive value of high-dimensional omics data over classical clinical predictors and/or assess the respective importance of the two types of predictors? Leaving the first challenge aside, we focus on tests and compare globalboosttest (Boulesteix and Hothorn, 2010) and pre-validation (Tibshirani and Efron, 2002), two testing approaches. Since omics data are high-dimensional, standard likelihood ratio tests in the framework of Generalized Linear Models (GLM) cannot be performed. The two examined methods tackle this problem based on a two-step procedure, but in different ways: while globalboosttest summarizes clinical predictors as an offset and then fits a regularized regression model to omics data, pre-validation first summarizes omics data as a cross-validated "pseudo predictor" and then tests its significance in a multivariate GLM adjusting for clinical predictors.

2 Globalboosttest

The "globalboosttest" procedure (Boulesteix and Hothorn, 2010) aims at testing the additional predictive value of high-dimensional data by combining two well-known statistical tools: GLMs and boosting regression. Suppose we have both high-dimensional omics data as potential predictors on the one hand and a few classical clinical covariates or a well-defined prognostic index on the other hand. The considered null-hypothesis is that "given the clinical covariates the omics data have no added predictive value". To address this testing problem the globalboosttest procedure first builds a clinical model (step 1, also denoted as "internal in this paper") based on clinical covariates only. For example step 1 is based on logistic regression in case of a binary response or on Cox proportional hazard regression in case of a censored survival response. The resulting linear predictor is then considered

as an offset in a more complex model involving omics data. As suggested by the procedure's name, the latter model is estimated by boosting regression (step 2, also denoted as external in this paper). Step 2 implies an iterative stepwise selection of the omics predictors while taking the clinical covariates into account in form of the offset. This step 2 is then repeated a large number of times after randomly permuting the omics data (but not the clinical data). A permutation p-value is then derived as the frequency at which the negative binomial log-likelihood of the boosting regression model was smaller in permuted data sets than with the original data set.
Tuning parameters are the number of boosting iterations in the external model and the number of permutations performed in step 2. The number of permutations should be as large as computationally feasible to increase the test's precision. The number of boosting regression steps is a parameter that potentially influences the test result.
Note that, strictly speaking, this permutation procedure tests the joint hypothesis that "omics data have no added predictive value" *and* "omics data and clinical predictors are independent", because by permuting omics data we also destroy the association between omics and clinical predictors. An important feature of the globalboosttest procedure, however, is that the offset is fixed and computed before seeing the omics predictors. Thus, in the case where omics and clinical predictors are strongly correlated, we expect the clinical offset to capture much variability and hence the null-hypothesis to be retained. This issue will be further discussed in Section 4.
The fact that the offset is fixed also implies that the coefficients of the clinical predictors fit in step 1 are *not* influenced by the omics predictors added to the model by boosting regression in step 2. On the one hand such an offset can well address the question of the *added* predictive value. The offset can be considered as an artificial but compulsory first predictor that is subsequently completed by the omics predictors selected afterwards. On the other hand the inconvenience is that clinical covariates cannot be tested – either individually or as a whole. The globalboosttest procedure allows to test the omics predictors only.

In principle any type of response variable can be analysed using globalboosttest provided that it can be accommodated into GLMs and boosting regression. This includes normally distributed, binary or censored responses. Furthermore, boosting regression may be essentially replaced by any regularized regression technique allowing an offset, e.g. the Lasso.

3 Pre-validation

The pre-validation method is based on a classical hypothesis testing framework within a GLM including the clinical predictors as well as a "pseudo-predictor" summarizing the omics predictors. This pseudo-predictor can be derived either at the link scale (which is preferred here in the context of survival analysis) or at the predictor scale. In principle all methods that can handle a large number of predictors can be used for this purpose, e.g. boosting regression or Lasso regression. In

this study boosting regression is considered for the sake of consistency with the globalboosttest procedure described in Section 3.

The obtained pseudo-predictor summarizing the omics data, however, should not be tested in a multivariate regression model based on the data set that were used for its construction. This approach would strongly favor omics data, because the pseudo-predictor constructed from high-dimensional would overfit the data set. To overcome this problem Tibshirani and Efron (2002) suggest "pre-validation". The term pre-validation refers to a cross-validation (CV) performed within the considered data set. At each CV iteration j, a pseudo-predictor is derived from the omics 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 pseudo-predictor value for each observation. This "pre-validated" pseudo-predictor 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. This pre-validation step is denoted as "internal".

Finally, a multivariate regression model (denoted as "external" model) is fitted using this pre-validated pseudo-predictor and the clinical predictors as predictors. The added predictive value is assessed by testing the significance of the regression coefficient of the pseudo-predictor. However, in a subsequent publication (Höfling and Tibshirani, 2008) this test is shown to be biased due to the violation of the i.i.d. assumption in the GLM. Höfling and Tibshirani (2008) address this bias through a permutation procedure which we also use here.

In contrast to globalboosttest, pre-validation considers clinical and omics predictors more symmetrically - in the sense that the coefficients of the clinical predictors are affected by the omics data, which is not the case in globalboosttest. If clinical and omics data are correlated, we thus expect both the clinical predictors and the omics-based pseudo-predictor to capture an important part of the variability. Another difference to globalboosttest is that the pseudo-predictor is computed differently for the K subsets, thus making computation more intensive and interpretation more difficult.

4 Simulation study

Both methods address the added value of high-dimensional omics data in the same data situation, but essentially ask different questions. While globalboosttest directly focuses on the added predictive value, the rationale behind pre-validation is of more symmetric nature. In this paper, their respective performance is examined in different simulation settings with sample size $n = 200$ and a censored time-to-event as response Y. The partially unobserved survival times T_i ($i = 1,\dots,200$) are generated from a Cox-Weibull model (Cox, 1972) similarly to Binder and Schumacher (2008). The cumulative density is given as $F(t) = 1 - \exp(-(\lambda(t) \cdot t/\alpha))$, where $\lambda(t)$ denotes the hazard rate. The survival times T_i ($i = 1,\dots,200$) are generated as $T_i = \frac{-\log(U_i)\cdot\alpha}{\lambda(t)}$, whereby U_i is drawn from the uniform distribution $U(0,1)$ and

Setting	Number of omics covariates	Correlation coefficient	informative omics covariates	Intersection informative/correlated omics covariates
I	1000	$\rho = 0$	0	$\{\varnothing\}$
II	1000	$\rho = 0$	20	$\{\varnothing\}$
III–VI	1000	$\rho = 0.2$	20	$\{\varnothing, 5, 10, 20\}$
VII–X	1000	$\rho = 0.8$	20	$\{\varnothing, 5, 10, 20\}$
XI: Many omics predictors	5000	$\rho = 0.2$	20	$\{10\}$
XII: Few omics predictors	20	$\rho = 0.2$	20	$\{20\}$
XIII: Many informative omics	1000	$\rho = 0.2$	200	$\{20\}$
XIV: Few informative omics	1000	$\rho = 0.2$	2	$\{\varnothing\}$
XV: Perfect corrlation i	1000	$\rho = 1$	20	$\{20\}$
XVI: Perfect corrlation ii	1000	$\rho = 1$	0	$\{\varnothing\}$

Table 1 Overview on simulation settings with a censored time-to-event as response, 5 clinical covariates and $n = 200$. If so each clinical covariate correlates to 10 omics covariates. For boosting $m_{stop} = (50, 100, 200, 500, 1000)$ iterations and for testing 1000 permutations are considered. For each setting globalboosttest/pre-validation are computed on 100 different data sets.

the shape parameter α is set to 1 in our study. Assuming proportional hazards, $\lambda(t)$ is modeled as $\lambda(t) = \lambda_0 \exp(\eta)$, whereby the baseline hazard rate λ_0 is set to 0.1 and η denotes an additive predictor. The follow-up times F_i are generated independently of the survival times in the same way as T_i but with a constant hazard rate $\lambda(t) = 0.1$ and $\alpha = 1$. Observations are censored if their follow-up time ends before the expected event such that about half of the observations are censored. Finally, the observation times Y_i are obtained as $Y_i = \min(F_i, T_i)$. Further, we generate five standard normal and mutually uncorrelated clinical predictors as well as 1000 standard normal omics predictors. Ten of these omics predictors are correlated with the first clinical predictor, 10 further omics predictors are correlated with the second clinical predictor, and so on, yielding a total of 50 omics predictors correlated with clinical predictors, whereby the correlation ρ is set either to $\rho = 0$ (no correlation), $\rho = 0.2$ (weak correlation) and $\rho = 0.8$ (strong correlation). The linear predictor η is defined as follows. The regression coefficients of the clinical predictors are chosen to mimick a realistic scenario with predictors of varying strengths: $\beta_{clinic} = (0, 0.5, 2, -1.5, -1)^T$. Out of the 1000 omics predictors, 20 have non-zero regression coefficients in the linear predictor η. The 20 coefficients are

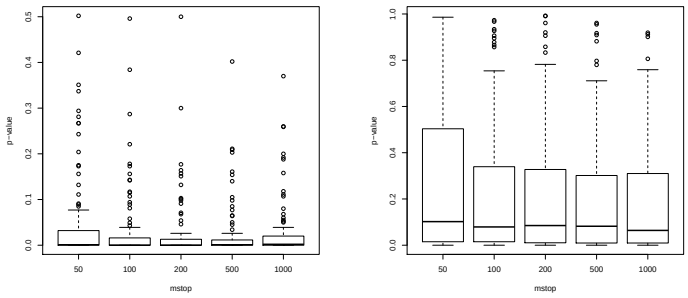

Fig. 1 Results of setting V. **Left:** globalboosttest. **Right:** pre-validation.

drawn from the uniform distribution $U(0.1, 0.7)$. Importantly, the size of the intersection between the 20 predictive predictors and the 50 correlated predictors is set successively to 0, 5, 10 and 20. Size 0 yields a setting where clinical and informative omics predictors are completely uncorrelated, while size 20 means that all informative omics predictors are correlated with clinical predictors. Table 1 (top) sums up the resulting settings, including an additional "null-scenario" without informative omics predictors (setting I). Moreover, some more extreme situations (bottom of Table 1) are additionally included in the study to complement the considered settings. First, setting VI is enlarged to 5000 omics covariates (setting XI) and as well reduced to only 20 (setting XII). Settings with more (setting XIII) and fewer (setting XIV) informative omics predictors are also considered. Finally settings with perfect correlation ($\rho = 1$) between the five clinical predictors and the 50 correlated omics predictors are considered, either with completely non-informative omics predictors (setting XV) or with 20 informative predictors included in the 50 correlated omics predictors (setting XVI). For each setting the two methods globalboosttest and pre-validation are evaluated based on 100 randomly generated data sets. For both globalboosttest and pre-validation the number of boosting iteration m_{stop} is set successively to 50, 100, 200, 500, and 1000. All tests base on 1000 permutations. As expected, both globalboosttest and pre-validation yield p-values that are approximately uniformly distributed $[0, 1]$ in the absence of informative omics predictors (data not shown). When omics predictors are informative and not perfectly correlated with clinical predictors, globalboosttest tends to yield smaller p-values than pre-validation. This result is illustrated by Figure 1 that displays the p-value of the two tests for different numbers m_{stop} of boosting iterations in Setting V. Moreover, globalboosttest tends to already reach good power for a smaller m_{stop} even in the case of high correlation between omics and clinical predictors. In contrast, pre-validation needs more iterations to capture the added predictive value of omics predictors. That is probably because pre-validation does not take the clinical predictors into account while summarizing the omics predictors and thus first captures information that are already captured by clinical predictors. By considering clinical predictors as an offset, globalboosttest captures the residual variability that is not captured by clinical predictors. Thus, globalboosttest generally needs less boosting iterations to reach good power. An exception is setting X, where *all* informative predictors are strongly correlated with a clinical predictor: globalboosttest then yields uniformly distributed p-values for a small m_{stop}, while a large m_{stop} leads to smaller p-values.This result obtained in setting X is related to the essential goal of globalboosttest. Globalboosttest tests the *added* predictive value of omics predictors and focuses on the part of the variability that is not captured by the clinical offset. In the unrealistic extreme case where all 50 informative omics predictors are perfectly correlated with a clinical predictor, the p-values are uniformly distributed on $[0, 1]$ with globalboosttest, but not with pre-validation. Note that this result is in contradiction with the theoretical null-hypothesis corresponding to globalboosttest: a strong correlation between omics and clinical predictors does not lead to the rejection of the null-hypothesis. Pre-validation employing the Lasso as "internal model" struggles with some problems related to tuning. As the number of observations is typically

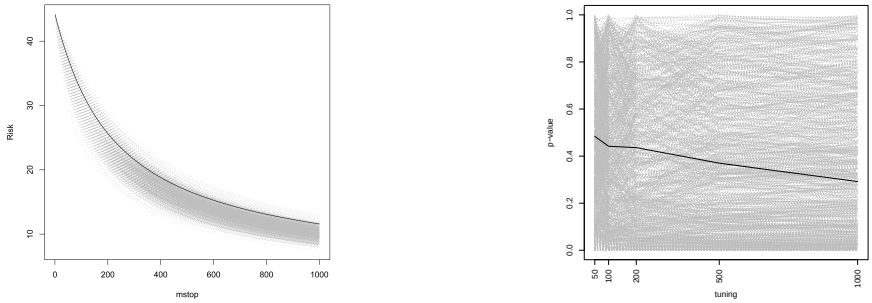

Fig. 2 Permutation procedure on Chin data. **Left:** globalboosttest. **Right:** pre-validation.

small for omics data, there are even less observations in training and test data sets. That makes the choice of λ extremely unstable. Each fold of the pseudo-predictor is based on a different value of λ. In many cases the optimal choice of λ selects no omics predictors at all. The choice of λ is much more crucial as the choice of boosting parameter m_{stop}. Consequently, globalboosttest and pre-validation employing boosting perform substantially better than pre-validation employing the Lasso.

5 Analysis of Breast Cancer Data

For illustration a breast cancer data set (Chin and et. al., 2006) including 77 patients is analyzed using globalboosttest and pre-validation with boosting regression. The response of interest is the censored distal recurrence time in years. The considered data set includes 11 clinical predictors such as age at diagnosis, variables of the TNM staging system or information on estrogen and progesterone receptors, as well as the expression level of 22215 genes acting as omics predictors.

The permutation-based p-values range from 0.77 to 0.97 for globalboosttest and from 0.29 to 0.48 for pre-validation (depending on m_{stop}). Figure 2 displays the curves representing the negative binomial log-likelihood (for globalboosttest) and the p-value (for pre-validation) obtained with the original data set (black) and the permuted data sets (grey). Both methods suggest that omics predictors do not improve prediction strength.

6 Summary and outlook

Simulation results suggest that in case of poor to moderate correlation between clinical and omics predictors globalboosttest tends to have a superior power to pre-

validation. However, in case of strong correlation globalboosttest becomes more conservative, which reflects its rationale: globalboosttest tests added predictive value, i.e. focuses on the variability that is not already captured by the clinical predictors. Whether it makes more sense to reject or the accept the null-hypothesis in the case of strong correlation depends on the substantive context. Correlation also seems to increase the impact of the number of boosting steps, suggesting that a systematic method for the choice of this parameter should be developed in the future.

In our paper the globalboosttest and pre-validation are assessed with respect to their performance as testing procedures. However, similar approaches may be adopted to derive combined prediction rules based on both clinical and omics predictors, see Boulesteix and Sauerbrei (2011) for an overview of such approaches. Due to its asymmetrical character giving more importance to clinical predictors, we expect the prediction rule derived from globalboosttest to perform poorly when these predictors are weak. A pre-validation approach may be promising, see Boulesteix et al (2008) for an example in the context of binary classification. More research is needed to assess the respective merits of the two methods in terms of predictive accuracy.

Acknowledgments

We thank Jutta Engel for helpful advice on the breast cancer data.

References

Binder H, Schumacher M (2008) Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. Statistical Applications in Genetics and Molecular Biology 7(1):12

Boulesteix A, Porzelius C, Daumer M (2008) Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. Bioinformatics 24(15):1698–1706

Boulesteix AL, Hothorn T (2010) Testing the additional predictive value of high-dimensional molecular data. BMC Bioinformatics 11:78

Boulesteix AL, Sauerbrei W (2011) Added predictive value of high-throughput molecular data to clinical data and its validation. Briefings in Bioinformatics 12(3):215–229

Chin K, et al (2006) Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. Cancer Cell 10(6):529–541

Cox DR (1972) Regression models and life-tables. Journal of the Royal Statistical Society, Series B 34(2):187–220

Höfling H, Tibshirani RJ (2008) A study of pre-validation. The Annals of Applied Statistics 2(2):643–664

Tibshirani RJ, Efron B (2002) Pre-validation and inference in microarrays. Statistical Applications in Genetics and Molecular Biology 1:1