Discussion of "The Evolution of Boosting Algorithms" and "Extending Statistical Boosting"

P. Bühlmann[1]; J. Gertheiss[2,3]; S. Hieke[4,5]; T. Kneib[6]; S. Ma[7]; M. Schumacher[4]; G. Tutz[8]; C.-Y. Wang[9]; Z. Wang[10]; A. Ziegler[11,12,13]

[1]Seminar for Statistics, Department of Mathematics, ETH Zürich, Switzerland; [2]Department of Animal Sciences, Biometrics & Bioinformatics Group, Georg-August-University of Göttingen, Göttingen, Germany; [3]Center for Statistics, Georg-August-University of Göttingen, Göttingen, Germany; [4]Institute for Medical Biometry and Statistics, Medical Center – University of Freiburg, Freiburg, Germany; [5]Freiburg Center of Data Analysis and Modelling, University of Freiburg, Freiburg, Germany; [6]Chair of Statistics, Georg-August-University of Göttingen, Göttingen, Germany; [7]Department of Biostatistics, Yale School of Public Health, Yale, USA; [8]Seminar of Applied Stochastics, Department of Statistics, Ludwig Maximilians University, München, Germany; [9]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, USA; [10]Department of Research , Connecticut Children's Medical Center, Hartford, USA; [11]Institute of Medical Biometry and Statistics, University of Lübeck, University Medical Center Schleswig-Holstein, Lübeck, Germany; [12]Center for Clinical Trials, University of Lübeck, Lübeck, Germany School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Keywords
Statistical computing, statistical models, algorithms, classification, machine learning

Summary
This article is part of a For-Discussion-Section of *Methods of Information in Medicine* about the papers "The Evolution of Boosting Algorithms – From Machine Learning to Statistical Modelling" [1] and "Extending Statistical Boosting – An Overview of Recent Methodological Developments" [2], written by Andreas Mayr and co-authors. It is introduced by an editorial. This article contains the combined commentaries invited to independently comment on the Mayr et al. papers. In subsequent issues the discussion can continue through letters to the editor.

Correspondence to:
See list of authors´ adresses at the end of the article.

Methods Inf Med 2014; 53: 436–445
doi: 10.3414/13100122
epub ahead of print: November14, 2014

With these comments on the papers "The Evolution of Boosting Algorithms – From Machine Learning to Statistical Modelling" [1] and "Extending Statistical Boosting – An Overview of Recent Methodological Developments" [2], written by Andreas Mayr and co-authors, the journal seeks to stimulate a broad discussion on boosting. An international group of experts has been invited by the editor of *Methods* to comment on the paper. Each of the invited commentaries forms one section of this paper.

1. Comment by P. Bühlmann

We congratulate the authors for two thoughtful and stimulating papers on boosting methods. They present an informative overview, with applications from biomedical research in mind: it is very useful, provides valuable guidance and can serve as a basis for further developments.

1.1 Additional Thoughts on Boosting

Much has happened since the inception of the first boosting algorithm which goes back to Schapire [3] and Freund and Schapire [4]. My understanding is that Breiman [5] was first to point out that AdaBoost is a functional gradient descent scheme: with this explanation, he contributed in a pioneering way to clarify and increase our understanding of AdaBoost, and to pave the road to "statistical" boosting. Friedman et al. [6] and Friedman [7] have further built on this gradient descent idea and brought in many additional "statistical" views, and Tutz and Binder [8] introduced the related nice concept of likelihood-based boosting. More theoretical results were established in the machine learning community considering the margin's point of view [9], and Bühlmann and Yu [10] proved a first statistical minimax rate result for L_2 Boosting in the context of nonparametric function estimation. All these developments are mentioned in the papers by the authors as well.

Componentwise L_2 Boosting is known in the signal processing literature as matching pursuit [11]. A major motivation for such an algorithm was its computational runtime: matching pursuit, or componentwise L_2 Boosting, is only evaluating inner products and scales very well for large datasets. The statistical motivation of further bias reduction by refitting residuals, exactly as in L_2 Boosting, has been recognized already by Tukey [12] who proposed "twicing". While twicing consists of two iterations only (hence the name "twiceing"), L_2 Boosting is a generalization to a finite number of iterations.

Another worthwhile connection can be made to numerical analysis. Breiman [5] also pointed out that boosting is a Gauss-Southwell algorithm, and it was realized later that L_2 Boosting amounts to the scheme of Landweber iterations for solving e.g. inverse problems, see for example Bis-

santz et al. [13]. Another relation which has never been much explored is when moving from gradient to conjugate gradient methods, see for example Lutz and Bühlmann [14]: since partial least squares can be seen as a conjugate gradient descent method [15, cf.], this view might open a new connection between boosting and partial least squares.

1.1.1 The Importance of Software

The "statistical" boosting algorithms are implemented in various R-packages, nicely described by the authors who made important contributions themselves to open source software. I emphasize the importance of software for statistical methodology and applications (a major motivation of our paper Bühlmann and Hothorn [16] was to provide good software): the R-packages mentioned by the authors are role models for excellent software which is user-friendly and at the same time flexible enough to allow for incorporation of user-specific features. The R-software environment together with Bioconductor are great resources for further development of excellent (boosting) software, e.g. for new problems with large-scale data.

1.2 Boosting and l_1-norm Penalization

The Lasso [17] has become extremely popular for estimation in high-dimensional models. It was a big surprise when Efron et al. [18] presented arguments for showing the strikingly close similarity of Lasso and componentwise L_2 Boosting in linear models. Bühlmann and Yu [19] prove exact equivalence of componentwise L_2 Boosting and the Lasso in an orthonormal linear model, and they also introduce "Sparse Boosting" which is equivalent to the adaptive Lasso [20] in an orthonormal linear model. Thus, componentwise boosting and sparse boosting, which are obviously sparse estimation techniques, can be viewed as some "sort of l_1-norm sparse" methods: although the connection is vague in general, I believe it allows for a useful interpretation.

The Lasso and l_1-norm regularization dominate nowadays the landscape in high-dimensional statistics: there are algorithms available which are mathematically justified, and there is a lot of detailed statistical theory [21, cf.]. Yet, boosting has advantages in terms of computational speed as well as adaptations to include additional constraints, e.g., monotonicity, by simple algorithmic adaptations. The statistical theory is much harder to derive though as one needs to analyze an algorithm [22, cf.] rather than the solution of a convex optimization problem.

1.3 Open Issues

When doing variable or feature selection in high-dimensional settings, a major obstacle is the high correlation or near linear dependence of a group of predictor variables. To cope with such a situation, Tutz and Ulbricht [23] propose a block-wise up-date in each boosting iteration. This is an interesting proposal but I wonder whether one needs more radical approaches, say when analyzing SNP data with $p \approx 10^6$ predictor variables. One idea would be to do the fitting on different levels of resolution [24, cf.]: here, the resolution is the size of a group of correlated variables and the groups could be constructed from hierarchical clustering. Bühlmann et al. [25] investigate some possible directions for estimation with groups of correlated variables, but further work is needed. In the context of boosting, the question arises whether one can construct a boosting algorithm in a suitable hierarchical fashion.

Another emerging theme is the analysis of large-scale ("big") data. Boosting algorithms, with their good computational scaling properties, are certainly interesting tools: however, I believe that one needs to account for potentially substantial heterogeneity in such large-scale data, see for example Meinshausen and Bühlmann [26].

As final words: addressing new important issues will help keeping boosting algorithms up to date.

2. Comment by J. Gertheiss

First of all, I would like to congratulate the authors on these two very well written papers illustrating the applicability and flexibility of the boosting concept when building statistical models. Hopefully these papers will stimulate the broader use of boosting for answering important research questions, for instance, but not only, in biomedicine. From my point of view this is the decisive next step boosting has to make to be broadly recognized as a substantially useful (statistical) method for data analysis. At the moment, however, boosting has not reached this stage, but is rather in danger of being left behind by alternative methods such as L_1-regularization.

To make one thing clear, I am not saying that L_1-regularization is the generally superior concept in terms of statistical methodology. Indeed, there are cases where one method is clearly preferable to the other. In [27], for example, L_1 would hardly be able to perform model selection, as is possible with boosting. By contrast, boosting cannot be used for fusion of categories of nominal predictors [28, 29]. In most situations, however, boosting and L_1-regularization are both applicable and closely related. In the high-dimensional linear model, for instance, there is the original lasso [17] and component-wise L_2-boosting [22]. In high-dimensional additive models, we can use mboost [16, 30–32] or L_1-regularization as proposed by [33]. With ordinal predictors, boosting and a grouped lasso-variant are both valuable approaches (see [34]); and for feature selection in signal regression we can use both block-wise boosting [35] and a structured elastic net [36], just to name a few. Nevertheless, when it comes to model and/or variable selection, most people seem to prefer the lasso or variants thereof. When looking at two of the most important, more recent papers about statistical boosting, [16] and [8], we see that these papers have been cited (according to the Web of Science on July 7, 2014) 134 and 46 times, respectively. Though these are remarkable numbers, two L_1-papers from the same years, [37] and [20], have been cited substantially more (600 and 801 times, respectively).

Apart from these pure numbers, which, as we all know, should not be overrated, there is another important aspect with respect to boosting that we should be aware of. There is a very active group of boosting enthusiasts contributing to the methodo-

logical development of boosting. This *boosting community*, as I will refer to them, has many members who know each other personally. Though this can be seen positively and is definitely a reason for the success of boosting in terms of methodological development, we have to keep it in mind when judging the "success" of boosting as a tool for practical data analysis. First, when we look at the more recent publications on boosting and citing articles, it has been my impression that many (or even most?) of these articles are written by authors that are part of the boosting community. Second, applied papers using boosting methods to answer research questions very often have coauthors from the boosting community. For instance, three out of the four applied papers [38–41] cited in the Conclusions of [2] even have coauthors that are also coauthors of [2].

My explanation for this is not that people outside the boosting community do not *like* boosting, but simply that they do not know enough about it. I have even met a number of statisticians who had never heard about boosting – but everyone knows the lasso (not everyone likes it but everyone knows it).

I am not saying that all this is bad news. It is perfectly normal for a relatively new method. But now it is time for boosting to make the next step, to spread beyond the core boosting community, and to become a "standard" tool for statistical model building and selection. Applied researchers should know that boosting is not so complicated to need boosting experts on board to be able to apply it. There is plenty of easy-to-use boosting software available, such as mboost.

To conclude, hopefully papers like the ones presented will drive more and more applied researchers to use boosting methods to answer their research questions. Only this will make it a *successful* statistical method. And you deserve it, boosting, you deserve it.

3. Comment by S. Hieke and M. Schumacher

In this issue of *Methods*, Mayr et al. [1, 2] provide comprehensive overviews on the evolution of boosting algorithms as well as on extending statistical boosting. The first addresses the fact that the roots of boosting can be located within the machine learning community. The current Wikipedia entry [42] starts with the following statement:

"Boosting is a machine learning meta-algorithm for reducing bias in supervised learning. Boosting is based on the question posed by Kearns [43]: Can a set of weak learners create a single strong learner? A weak learner is defined to be a classifier which is only slightly correlated with the true classification (it can label examples better than random guessing). In contrast, a strong learner is a classifier that is arbitrarily well-correlated with the true classification. Schapire's affirmative answer [3] to Kearn's question has had significant ramifications in machine learning and statistics, most notably leading to the development of boosting."

The early applications of boosting were thus focused on classification problems. One example published in this journal was a comparative investigation, improvement and evaluation of record linkage methods by Sariyar et al. [44]. Another example by Stollhoff et al. [45] aimed to evaluate various boosting variants in comparison to standard logistic regression for differentiating between benign and malignant breast tumors [46]. In a recently published article, Liu et al. [47] use adaptive boosting for improving the classification of elderly patients at high risk for drug-to-drug interactions. With the extension of boosting towards regression problems, a realm of statistical modelling, the potential field of applications is now much wider, and it is sometimes not straightforward – at least not for the non-expert – to recognize the original boosting idea as the core of a sophisticated regularized regression modelling approach, e.g. for high-dimensional data.

Insofar, the two papers [1, 2] in this issue are most welcome; they nicely complement the contributions to the focus theme "Recent development in boosting methodology" published about three years ago [48].

With the many faces of various boosting approaches the range of applications became multifaceted as well. In the following, we briefly describe a successful but rather unusual application for estimating the comprehensiveness of literature searches for systematic reviews [41]. In systematic reviews of clinical trials it is crucial that all relevant studies are identified through systematic literature searches in order to be included into the systematic review [49]. Therefore current advice is to base the search strategy on a number of relevant databases. For estimating the number of missing references, so-called capture-recapture techniques have been proposed. Most readers will know these techniques from the task to estimate the number of fish in a pond. For doing so, a first sample of fish is drawn and marked before returned into the pond. Afterwards, a second independent sample is drawn and from the relation of marked fish to all sampled fish the total number is inferred. This technique has been further extended to the search in different databases where from the relation of articles found in different databases to all articles found, the number of missing articles is estimated. In doing so the crucial point is that the statistical model used for estimation involves a large number of high-order interaction terms and it is unclear how to select the relevant interaction terms best without running the risk of overfitting. In contrast to manual selection, application of GAM Boost [8, 50] proved to be able to create an appropriate model for inference where a reliable estimate of the number of missing studies could be based on.

This example underlines that boosting is not only valuable in classification problems and in regression models with high-dimensional predictor variables but also in statistical models with a few predictors but with many additional structural parameters, in this case higher-order interaction terms that have to be taken into account.

The review papers [1, 2] also discuss different versions of boosting techniques (statistical boosting), e.g. gradient boosting and likelihood-based boosting sharing similar structures. Both statistical boosting algorithms can deal with "large p, small n" situations, which is specific in high-dimensional settings. Recently, a lot of effort was invested in the extension of the boosting concept from the machine learning community, i.e. AdaBoost algorithm for clas-

sification problems, towards almost any type of regression problems including time-to-event data. In addition to the relevant property of statistical boosting regarding variable selection, the developments concerning automated model choice, i.e. different types of predictor effects allowing linear and non-linear predictor effects on response, provide a flexible framework for statistical boosting. Given these methodological extensions and the implementation of statistical boosting algorithms in freely available open source R add-on packages together with the possibility of parallel computing reducing the computational burden, the concept of boosting becomes a flexible tool in biomedical research. Therefore, statistical boosting has been made available for a wide area of applications including application in high-dimensional molecular data settings where there are more predictor variables than observations. Besides the practical feasibility of statistical boosting for molecular data, it can be expected that statistical boosting algorithms can be used not only for single high-dimensional molecular data, but also for more complex situations such as the integration of different data sets from various molecular levels into a risk prediction model including clinical predictors as mandatory where the outcome variable can be continuous or represent time-to-event.

4. Comment by T. Kneib

4.1 Introduction

I really enjoyed reading this excellent overview on the evolution of statistical boosting from the machine learning origins and the current state of the art in statistical boosting. Hopefully, both papers will be helpful in convincing many interested scientists working on medical applications that boosting can be a useful tool for their research. In particular, presenting both likelihood-based boosting and functional gradient descent boosting jointly is certainly useful and allows to fully appreciate the advantages of boosting. Most notably, and as highlighted in the papers under discussion, boosting provides seamless integration of model choice and variable selection within the estimation of complex regression relationships.

While automated variable selection and model choice are provided by both functional gradient descent boosting and likelihood-based boosting, several open problems deserve further attention in future research (as acknowledged in the papers under discussion). I will comment on some issues that I consider to be of particular relevance below.

4.2 Modularity of Boosting

In addition to the automated variable selection, the main advantage of boosting is its modularity. While not reaching the same level as for instance Markov Chain Monte Carlo simulation techniques where complex, hierarchical models can be split into small pieces in a divide and conquer strategy, functional gradient descent boosting allows for a clear separation between the loss criterion describing the estimation problems (and therefore determining the working observations) on the one hand and the construction of suitable baselearners to implement a certain model structure on the other hand. This also allows for a very modular implementation as provided by the R add-on package mboost where new loss types can easily be implemented (and the same is true for baselearners although the internal structure is somewhat more complex in this case). This seems to be much more challenging in likelihood-based boosting. As an example, consider quantile boosting where a new optimisation procedure would have to be considered to minimize the check function. Even if standard optimizers are available, their combination with a new type of (e.g. monotonicity constrained) baselearner requires redeveloping major parts of the methodology and implementation.

4.3 Model Complexity and Variable Importance

For a proper evaluation of a model obtained with boosting, it is useful to provide additional information concerning the complexity and importance of the additive model components corresponding to the different baselearners. While the papers under discussion provide several comments in this direction, this point still seems to be largely unsolved or the proposed procedures are of a relatively large computational cost. Simply including any covariate in the final model that has been selected at least once until the optimal stopping iteration typically leads to many false positive detections and many baselearners with a rather weak contribution to the overall predictor. The frequency of selections is unfortunately not a suitable criterion since the importance of a predictor component also crucially depends on the iteration index when it has been included with early inclusions usually inducing more important contributions. Two potentially useful measures could be i) the norm of the contributions of a baselearner to the overall predictor, e.g.

$$\left\| \hat{h}_j(\cdot) \right\| = \sqrt{\sum_{i=1}^{n} \left(\hat{h}_j^{[m_{\text{stop}}]}(x_{ij}) \right)^2} \qquad (1)$$

where rarely selected effects should show up only with minor contributions and ii) the accumulated reduction in the model fit criterion provided by one baselearner

$$\text{Importance}\left(\hat{h}_j(\cdot)\right) = \sum_{m:\, j_m^* = j} \left(\rho\left(y, \hat{f}^{[m]}\right) - \rho\left(y, \hat{f}^{[m-1]}\right) \right) \qquad (2)$$

where j^*_m denotes the best fitting baselearner in iteration m. One issue with the norm (1) is that the total variability of the predictor unfortunately does not additively decompose into the sum of the individual contributions. For the baselearner importance measure (2), it would have to be ensured that indeed the fit criterion monotonically decreases with the boosting iterations such that

$$\rho\left(y, \hat{f}^{[m_{\text{stop}}]}\right) - \rho\left(y, \hat{f}^{[0]}\right) = \sum_{j=1}^{p} \text{Importance}\left(\hat{h}_j(\cdot)\right).$$

4.4 Unbiased Model Selection

For the comparison of baselearners with different flexibility, it is highlighted in the

papers under consideration that it is important to make them comparably in terms of their degrees of freedom. While it is not much of a surprise that I agree, I would still consider this to be a partially unsolved issue. In particular, comparing baselearners corresponding to a very high-dimensional effect (such as individual-specific or spatial effects) with a single parametric coefficient seems problematic since, albeit the comparable degrees of freedom, the complexity of the effect is distributed across many parameters in one case and completely assigned to one parameter in the other case. In my experience, boosting then tends to select the high-dimensional effect too rarely. This may be related to the fact that Hofner et al. [51] only studied the expected reduction in L_2-loss and not the complete distribution or other types of loss functions (which admittedly will be much more difficult). An issue also related to the selection and comparison of effects is given when baselearners are highly colinear. This may, for example, be the case for spatial effects and spatially varying covariates. In some cases, constructing orthogonal baselearners may partially resolve this issue but this will only be possible if effects have some kind of natural ordering. This works for example in case of a polynomial model where orthonormal polynomials can be used to define orthogonal baselearners for the separate coefficients of the polynomial. These make the l-th order polynomial orthogonal to the $(l-1)$ baselearners for the lower order polynomial contributions. This is in contrast to spatial effects and spatially varying covariates where it is not automatically clear whether the covariates should be orthogonalized with respect to the spatial effect or vice versa.

4.5 Boosting for Low-dimensional Models

My final question concerns the suitability of boosting for determining low-dimensional models. While formally the automated variable selection property should still hold in such models, my experience seems to suggest that boosting tends to select too many (if not all) covariates in such situations.

Acknowledgments

Financial support by the German Research Foundation (DFG), grant KN 922/4–1/2 is gratefully acknowledged.

5. Comment by S. Ma

In two consecutive papers [1, 2], the authors, Drs. Mayr, Binder, Gefeller, and Schmid, provided a comprehensive and timely review of the history, new developments, and applications of boosting methods. In the literature, there have been a large number of methodology, application, and review papers and books on boosting, many of which are referred in these two review papers. Yet, the present papers distinguish themselves and advance from the literature in many different ways. The authors should be applauded for their work.

The authors first provided a brief but still comprehensive review of the history of boosting. The evolvement of boosting can shed light on the directions of future development. Different from many published studies, the present review provides the intuition and rationale beneath the development of Adaboost and the more recent statistical boosting. Another major contribution that must be highlighted is a clear description of the distinction and connection between gradient boosting and likelihood-based boosting. The boundary between gradient boosting and likelihood-based boosting gets blurred in recent studies. Yet I believe the present review is among the first to unify them under the same statistical framework. Another feature, which makes this review especially suitable for the readers of MIM and beyond, is the emphasis on biomedical applications. The analysis of "classic" low-dimensional biomedical data using boosting has been extensively discussed in the literature. In comparison, the fast-moving high-dimensional field deserves more attention. The present review provides a comprehensive list of recent high-dimensional data analysis using boosting. With limited space, some details are missing. But the readers should have no trouble locating the original studies if needed.

In multiple occasions, the authors made direct connections between the boosting and penalization techniques. Under special settings (see for example those in [19]), the boosting and penalization estimates coincide. Under more general settings, although both boosting and penalization have the shrinkage estimation and selection properties, a comprehensive numerical and methodological comparison is still lacking. It was reviewed by the authors and is worth emphasizing that the boosting technique has also been used to compute penalized estimates (especially Lasso based) and played an important role before the coordinate/gradient descent techniques became popular [52]. Our limited numerical experience suggests that, in comparison to coordinate and gradient descent, the boosting-based algorithms for penalization can "get close" to the optimizer faster in the initial iterations, however, may take much longer to "get closer" (converge). Studies such as [52] provided some justification on the validity of boosting algorithms.

As partly reflected in the present review, it is interesting to try to "match" the history of boosting with that of penalization. Individual-variable-based penalization, group penalization, and hierarchical penalization can all easily find their boosting counterparts. For the analysis of high-dimensional biomedical – for example genetic – data, boosting methods have been developed to account for the pathway structure and connections between variables. The intuition behind such methods and corresponding references have been provided in this review. Under the penalization framework, methods with a smoothing effect, such as the fused Lasso and the more recent Laplacian penalization [53], have been proposed to accommodate finer data structure. The intuition is to promote the similarity in regression coefficients of physically (or statistically, biologically) adjacent variables. It will be of interest to further develop boosting methods that have a similar function.

As has been reviewed, boosting is believed to be "less susceptible" to over-fitting. Thus an old wisdom is to run boosting for a large number of iterations. However, under the high-dimensional setting where variable selection can be as important as estimation, one needs to be more

careful with the stopping rule. Bühlmann and Yu [19] and others recognized the possible over-selection of regular boosting and developed the sparse boosting and other methods. The over-selection problem is at least partly caused by how a variable is defined as "selected". Under the most extensively adopted estimation-based selection, a variable is selected as long as its estimate is nonzero, no matter how small or insignificant it is. In addition, with the connection between boosting and penalization, it is suspected that correlation among variables may also contribute to the over-selection. As has been reviewed, stability- and inference-based selection have been developed as generic ways of selection, have been applied to penalization and other techniques, and are potentially applicable to the existing boosting methods to tackle the over-selection problem. It is interesting to take another look at sparse boosting. With a step size less than one, the standard boosting technique already has a shrinkage property. With sparse boosting, a penalty is explicitly added in the selection of weak learners and stopping. There are multiple choices for this penalty, including not only AIC/BIC but also penalties that more heavily depend on the magnitudes of coefficients (such as Lasso type). It is unclear, at least to me, whether sparse boosting has a double-shrinkage problem, and whether a relaxation technique (in a similar spirit as twin boosting) can be applied to further improve selection and estimation.

As the authors pointed out, boosting has demonstrated significant merit for high-dimensional biomedical data in terms of selection, estimation, and prediction. It has been applied to the analysis of continuous traits, categorical disease status, and prognosis outcomes under various statistical models. The authors suggested, which I fully agree, that the popularity of boosting in data analysis is due to its intuitive formulation, simplicity in programming, easy adaption to different models, and availability of public software packages. We have applied boosting to integrative analysis, which has one more dimension than "standard" high-dimensional data analysis. In [54], we developed an integrative sparse boosting method, which as can be seen from the name is built on [19], and collectively analyzed multiple independent datasets. We showed that integrative sparse boosting can be more effective than "individual-dataset boosting + meta-analysis". Another type of integrative analysis, which has gained a lot of attention in the recent literature, is on data with multiple types of high-dimensional measurements on the same subjects. To the best of our knowledge, boosting methods for such data remain to be developed.

Boosting has played an important role in "classic" machine learning. It keeps attracting attention in recent machine learning and statistics research, as partly reflected in the present review and a recent issue of MIM dedicated to boosting. Its value for biomedical data analysis cannot be over-emphasized. The authors have conducted an outstanding review and also paved road for future development. I expect that this review will sparkle more interest and discussions among the readers of MIM and beyond, for methodological development and biomedical applications. Once again, I congratulate the authors for their effort and contribution.

6. Comment by G. Tutz

The authors are to be congratulated to a lucid presentation of the evolution of boosting concepts and current boosting methodology. The methods considered range from the early AdaBoost algorithm to the statistically motivated steepest gradient descent methods and likelihood-based boosting algorithms. When presenting the methods the authors focus on the (generalized) additive model with selection referring to the additive components. They also refer to extensions to other settings like survival models but the algorithms given are restricted to the selection of additive components. Therefore, in the first comment we aim at giving a more general framework for boosting.

6.1 Boosting for a General Class of Models

Boosting can be seen as a very general regularization method for structured regression that is able to simultaneously estimate and select interesting features in the predictor space. The features can be components in an additive model but also interaction terms, varying coefficients or whole vectors of dummy variables. It seems worthwhile to give a more generic form of the algorithm that covers more interesting cases. In the following we consider a more general boosting algorithm in the spirit of likelihood boosting but not restricted to classical likelihood based boosting for additive models.

Let us consider data (y_i, x_i), $i = 1, \ldots, n$, where y_i is the response and $x_i^T = (x_{i1}, \ldots, x_{ip})$ is a vector of explanatory variables. A general class of models specifies that the mean $\mu_i = E(y_i|x_i)$ is determined by $\mu_i = h(\eta_i)$, where $h(.)$ is a known response function and η_i contains the explanatory variables in a structured form. In addition, one assumes that $y_i|x_i$ follows a specific distribution, often a distribution from the simple exponential family. This class of models includes univariate generalized linear models (GLMs) with univariate response y_i and mean $\mu_i = h(\eta_i)$, where $\eta_i = x_i^T\beta$, but also generalized additive models (GAMs), which assume $x_i = h_1(x_{i1}) + \ldots + h_p(x_{ip})$. Also models with effect modifiers can be given in this form. A general boosting algorithm for this class of models is the following.

Structured Regression Boosting

Step 1 (Initialization)

For given data (y_i, x_i), $i = 1, \ldots, n$, fit an initial model, typically the intercept model $\mu^{(0)}(x) = h(\beta_0)$ to obtain an estimate $\hat{\eta}^{(0)} = \hat{\beta}_0$.

Define parametrically structured terms , $\eta_j\ (x, y_j) = 1, \ldots, m$ that serve as base learners. Typically starting values are $\hat{y}_j^{(0)} = 0$

Step 2 (Iteration) For $l = 0, 1, 2,...$

1. *Estimation step:* Fit the models

$$\mu_i = h\left(\hat{\eta}^{(l)}(x_i) + \eta_j(x_i, \gamma_j)\right),$$

$j = 1, \ldots, m$ to data (y_i, x_i), $i = 1, \ldots, n$, where $\hat{\eta}^{(l)}(x_i)$ is treated as an offset and the predictor is estimated by fitting the parametrically structured term $\eta_j(x_i, y_j)$, obtaining $\hat{y}_j$.

2. *Selection step:* Select the structured term $\eta_{j^*}(x_i, \hat{y}_{j^*})$ that showed the best performance
3. *Update:* The improved fit is obtained by $\hat{\eta}^{(l+1)} = \hat{\eta}^{(l)}(x_i) + \eta_{j^*}(x_i, \hat{y}_{j^*})$, $\hat{\mu}_i^{(l+1)} = \mathrm{h}(\hat{\eta}^{(l+1)}(x_i))$
   The improved parameters are obtained by

$$\hat{\gamma}_{j^*}^{(l+1)} = \hat{\gamma}_{j^*}^{(l)} + \hat{\gamma}_{j^*}.$$

The strength of the algorithm is in the definition of the base learners, which can be chosen with reference to the structure that is interesting and to be selected and fitted. In particular the base learner can contain a set of explanatory variables yielding blockwise boosting methods. Some examples are

- $\eta(x_i, y) = x_{ir}y_r$ which specifies the linear effect of the rth covariate;
- $\eta(x_i, y) = y_0 + x_{ir}y_r$, which specifies the intercept and the linear effect of the rth covariate;
- $\eta(x_i, y) = x^T_{ir}y_r$, where x_{ir} is a vector of dummy variables corresponding to a categorical variable (blockwise boosting);
- $\eta(x_i, y) = x_{ir}x_{is}y_{rs}$, representing an interaction between the rth and the sth covariates;
- $\eta(x_i, y) = x_{ir}x^T_{is}y_{rs}$, representing an interaction between the rth variable and the sth categorical variable given by a vector of dummy variables;
- $\eta(x_i, y) = \sum_j \gamma_j B_j(x_{ir})$, where $B_1(.)$, $B_2(.)$, ... are basis functions, for example, B-splines; the base learner represents a smooth function of the rth variable
- $\eta(x_i, y) = x_{is}\left(\sum_j \gamma_j B_j(x_{ir})\right)$ representing that the effect of variable x_{is} varies smoothly over variable x_{ir}.

Thus the set of chosen base learners defines the possible structures or combination of structures that are fitted. The fit itself can be obtained by maximization of a (possibly penalized) likelihood or by using weighted least squares estimates motivated by gradient descent as in L2 boosting. The crucial point is that the estimate improves the fit only slightly to obtain a weak learner.

6.1 Extensions to Multivariate Settings

The same basic algorithm can be used for multivariate responses. Then data have the form (y_i, x_i), $i = 1, ..., n$, where y_i is a q-dimensional vector-valued response and the mean is a vector $\mu_i = \mathrm{E}(\mathrm{y_i}|\mathrm{x_i})$ determined by $\mu_i = h(\eta_i)$, where $h(.)$ is a q-dimensional response function and also η_i is a vector. In particular, multivariate GLMs for multinomially distributed y_i fit into this framework, see Fahrmeir and Tutz [55] or Tutz [56]. Boosting methods for the multinomial logit model and for ordinal response models, which use this form have been proposed by Zahid and Tutz [57] and Zahid and Tutz [58]. The additional challenge in multinomial models is that the effect of each explanatory variable is given by a set of parameters because one needs one parameter for each response category. Therefore, to obtain variable selection one has to collect the whole set of parameters in the base learner such that they are simultaneously included or not. This is a form of blockwise boosting tailored to multinomial responses. The other multivariate structure that is interesting are generalized mixed model for clustered data, which can also be embedded into this framework by letting y_i contain the repeated measurements on the ith cluster (see Groll and Tutz [59]).

6.2 Boosting versus Regularization by Penalties

Selection of relevant features in structured regression can be obtained by structured boosting methods or, alternatively, by penalized maximum likelihood estimation. With the seminal paper of Tibshirani [17] the lasso and its various extensions have become intensively used regularization methods in a wide range of areas. It is known that for the selection of variables in simple linear models boosting with very weak learners yields similar results as the lasso. We found, considering more general settings, that when selecting ordinal predictors boosting selects fewer variables than methods based on fusion penalties (Gertheiss et al. [34]). But a thorough comparison of boosting and penalty methods and an evaluation of the pros and cons seems to be missing. It would be interesting to compare the methods more systematically because they are competitors in many selection problems but are quite different in their construction. While boosting is an algorithmic regularization technique that obtains selection by early stopping penalty methods use an explicit penalty term. It shows which parameters or features are penalized and therefore shrunk toward a specific value, which may be seen as an advantage of the method. In some cases it seems to be simpler to define a penalty than to find a corresponding boosting method. For example, in discrete structures, if one wants to find clusters of a categorical variable fusion penalties as considered by Tutz and Gertheiss [60] are an effective tool to identify categories that share the same effect. Corresponding boosting methods seem not to be available because weak learners that find clusters are hard to construct.

7. Comment by Z. Wang and C.-Y. Wang

We congratulate the authors (MBGS) for presenting a rather comprehensive review on the boosting technology. They nicely summarize broad applications of boosting methodology in many statistical problems. The illustrations of open source R packages can help other researchers in their own works.

In this discussion we would like to comment on loss function, gradient boosting vs likelihood gradient, significance level and boosting, and applications of boosting in missing data.

7.1 Loss Function

MBGS state that "The gradient boosting approach can be used to optimize any loss function that is at least convex and differentiable". Most of the loss functions described in MBGS are convex and differentiable. However, several loss functions deserve special consideration since they are not convex and differentiable everywhere but still applicable with boosting. Consider

the hinge loss for a classification rule f given binary outcome $y \in \{-1, 1\}$:

$$(1 - yf)_+,$$

where $z_+ = \max(0, z)$. The hinge loss is not differentiable at $f = y$, however, this event has probability 0 to be realized by the data. HingeBoost was developed based on the hinge loss [61]. Similar examples include the multi-class hinge loss [62], the absolute loss [16] and additive quantile regression [27].

7.2 Gradient Boosting vs Likelihood Boosting

MBGS suggest a unified framework for gradient boosting and likelihood-based boosting. We believe the gradient boosting is more general, not only because it can be applied in problems for which likelihood is not defined, but there are significant differences between the two algorithms. The likelihood-based boosting requires computing the largest log-likelihood in step (4) for a given distribution. Therefore, one may have to rely on some well developed algorithms for this task. This is not a problem for some well studied models, such as the exponential family distributions. However, the gradient boosting is a stand-alone algorithm, and can innovatively solve new problems. For instance, to make HingeBoost in the framework of likelihood-based boosting, it seems one may have to rely on HingeBoost itself or its cousin support vector machine [63].

7.3 p-Values for Boosting

As demonstrated in MBGS, boosting is a powerful tool for variable selection in a wild range of data. However, results from applying boosting may still contain noise variables, particularly in high-dimensional regression. In biomedical research, it is important to distinguish between effective and non-effective variables and applied medical researchers are customized with the notion of "p-value". While significance level in high-dimensional regression is still an active research topic, the multi-split method has been proposed for methods including boosting [64]. In this method, the data are randomly and disjointedly split to equally sized selection and p-value sets. We apply the boosting algorithm to the selection data set to determine effective predictor variables. The traditional regression is then applied to the p-value data set with only those effective variables. Hence we obtain p-values for the selected variables. On the other hand, the p-values are 1s for the non-selected variables. Next, the p-values are adjusted, for instance, with Bonferroni technique. The above procedure is repeated for B times and we have a total of B p-values $P_{j,b}$ for each predictor $j = 1, \ldots, p$, $b = 1, \ldots, B$. For each $j = 1, \ldots, p$, the following summary statistics can be used:

$$Q_j(\gamma) = \min\{q_\gamma(P_{j,b}/\gamma;\, b = 1, \ldots, B), 1\},$$

where $\gamma \in (0, 1)$ and $q_\gamma(\cdot)$ is the empirical γ-quantile function. A p-value is given by $Q_j(\gamma)$ for each predictor variable $j = 1, \ldots, p$, for any fixed $0 < \gamma < 1$. This value is an asymptotically correct p-value for controlling the familywise error rate for the L_2 boosting under appropriate conditions. Furthermore, an optimal value of γ can be obtained as in [64]. We illustrate how to compute p-values with two examples in the online supplementary material.

7.4 Imputation for Missing Predictors in Boosting

Imputation for missing predictors in boosting is an important research topic in medical research. To this problem, Wang and Zeng [65] proposed imputation methods in boosting. In general, if the purpose is prediction, then the conditional mean imputation methods in [65] would be valid. However, if we are interested in confidence interval estimation of the prediction when predictors may be missing, then additional work is required to take into account uncertainty due to imputation. Multiple imputation may be a valid approach to address confidence interval estimation with missing predictors. However, complications may arise due to high dimensional data with multiple imputations. Further investigation in this research is warranted.

Acknowledgments

ZW is partially supported by a grant from the Charles H. Hood Foundation, Inc., Boston, MA. CYW is partially supported by National Institutes of Health grants CA53996, R01ES017030 and HL121347.

8. Comment by A. Ziegler

Mayr et al. [1, 2] have to be congratulated for their excellent articles in which they explained boosting in a very simple way. The first paper [1] thereby closes an important gap in the literature. It starts with the classical AdaBoost, which is known to many researchers, then moves to the statistical way of boosting, i.e., gradient boosting and the more recent likelihood-based boosting. The authors establish the links between these approaches in clever ways. However, with more pages available, it would have been great to see even more links with other work related to boosting. For example, a step in-between AdaBoost and gradient boosting is Tukey's twicing [66]. In detail, Tukey proposed to run a linear regression twice. The first run was on the original data, the second one used the ordinary residuals obtained from the initial regression. The parallel here with boosting is that in a linear model the residuals form a weighted version of the original data.

Despite the excellent treatment of the boosting methods by the authors, several questions have not been addressed in the two review articles. One important aspect of any learning machine is whether the learning machine is consistent. A consistent estimator is defined to be any estimator for which the estimated quantity converges in probability to the true quantity [67]. In the classification context of boosting, an estimator is consistent if the classification rule converges in probability to the Bayes rule. As discussed by Mease and Wyner [67], AdaBoost is most likely not consistent as long as no regularization is employed. Another important question is the rate of convergence of the learning machine. A final important aspect which the authors discussed in detail is the optimal stopping iteration m_{stop} of the boosting algorithm. The authors criticized the use of standard

information criteria [1], such as Akaike's information criterion. They furthermore considered the use of resampling or cross-validation for tuning m_{stop}. This tuning of the optimal stopping rule performs well only in those cases where the model is generalized to data with the same structure, and one example, where this strategy might fail is the prediction of endpoints in stroke [68].

Boosting stumps was compared with bootstrap averaging of stumps or even larger classification trees in the literature several times, and a brief discussion can be found in Mayr et al. [1], Section 2.2. Breiman [69] extended bagging trees to random forests, and the most interesting component of this random forests is the random selection of features at each split in a tree. This random feature selection does not only integrate additional variability to a tree. It also is a way to cope with collinearity. With the additional property of random forests that the importance of a variable can be determined, e.g., through permutation, even two highly correlated features can both be identified as being important [70]. The random feature selection component in the tree building stage of random forests has another positive aspect. Depending on the proportion of features available at a split, generally denoted by *mtry*, the features are most likely uncorrelated. Currently, this random feature component can only be integrated through an external bootstrap loop in boosting. Such an extra resampling step has been nicely described by Mayr et al. [1] in their supplement. It would be great if these beneficial properties of random forests could be integrated into boosting algorithms.

Conflicts of interest

The author is member of the editorial board of the *Biometrical Journal*, *Methods of Information in Medicine* and *Statistics in Medicine*.

References

1. Mayr A, Binder H, Gefeller O, Schmid M. The Evolution of Boosting Algorithms – From Machine Learning to Statistical Modelling. Methods Inf Med 2014; 53 (6): 419–427.
2. Mayr A, Binder H, Gefeller O, Schmid M. Extending Statistical Boosting – An Overview Of Recent Methodological Developments. Methods Inf Med 2014; 53 (6): 428–435.
3. Schapire RE. The strength of weak learnability. Machine Learning 1990; 5 (2): 197–227.
4. Freund Y. Schapire RE. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1996. pp 148–156.
5. Breiman L. Prediction games and arcing algorithms. Neural Comput 1999; 11 (7): 1493–1517.
6. Friedman JH, et al. . Additive logistic regression: a statistical view of boosting (with discussion). Annals of Statistics 2000; 28 (2): 337–407.
7. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics 2001; 29: 1189–1232.
8. Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. Biometrics 2006; 62 (4): 961–971.
9. Schapire RE, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. Annals of Statistics, pages 1998. pp 1651–1686,
10. Bühlmann P, Yu B. Boosting with the L2 loss: regression and classification. Journal of the American Statistical Association 2003; 98: 324–339.
11. Mallat S, Zhang Z. Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing 1993; 41: 3397–3415.
12. Tukey JW. Exploratory Data Analysis. Addison-Wesley, 1977.
13. Bissantz N, et al. Convergence rates of general regularization methods for statistical inverse problems and applications. SIAM Journal of Numerical Analysis 2007; 45: 2610–2636.
14. Lutz RW, Bühlmann P. Conjugate direction boosting. Journal of Computational and Graphical Statistics 2006; 15 (2): 287–311.
15. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer, New York; 2001.
16. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting (with discussion). Statistical Science 2007; 22: 477–505.
17. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B 1996; 58: 267–288.
18. Efron B, et al. Least angle regression (with discussion). Annals of Statistics 2004; 32: 407–451.
19. Bühlmann P, Yu B. Sparse boosting. Journal of Machine Learning Research 2006; 7: 1001–1024.
20. Zou H. The adaptive Lasso and its oracle properties. Journal of the American Statistical Association 2006; 101: 1418–1429.
21. Bühlmann P, van de Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Verlag; 2011.
22. Bühlmann P. Boosting for high-dimensional linear models. Annals of Statistics 2006; 34: 559–583.
23. Tutz G, Ulbricht ,J. Penalized regression with correlation-based penalty. Statistics and Computing. 2009; 19 (3): 239–253.
24. Lee AB, et al. Treelets: an adaptive multi-scale basis for sparse unordered data (with discussion). Annals of Applied Statistics 2008. pp 435–500.
25. Bühlmann P, Rütimann P, van de Geer S, Zhang CH. Correlated variables in regression: clustering and sparse estimation (with discussion). Journal of Statistical Planning and Inference 2013; 143 (11): 1835–1871.
26. Meinshausen N, Bühlmann P. Maximin effects in inhomogeneous large-scale data, 2014. Preprint arXiv:1406.0596.
27. Fenske N, et al. Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. Journal of the American Statistical Association 2011; 106: 494–510.
28. Bondell HD, Reich BJ. Simultaneous factor selection and collapsing levels in ANOVA. Biometrics. 2009; 65: 169–177.
29. Gertheiss J, Tutz G. Sparse modeling of categorial explanatory variables. The Annals of Applied Statistics 2010; 4: 2150–2180.
30. Hofner B, Mayr A, Robinzonov N, Schmid M. A hands-on tutorial using the R package mboost. Computational Statistics 2014; 29: 3–35.
31. Schmid M, Hothorn T. Boosting additive models using componentwise P-splines. Computational Statistics & Data Analysis 2008; 53: 298–311.
32. Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B. mboost: Model-Based Boosting, 2014. R package version 2.3-0.
33. Meier L, Van de Geer S, Bühlmann P. High-dimensional additive modeling. The Annals of Statistics 2009; 37: 3779–3821.
34. Gertheiss J, Hogger S, Oberhauser C, Tutz G. Selection of ordinally scaled independent variables with applications to international classification of functioning core sets. Journal of the Royal Statistical Society C (Applied Statistics) 2011; 60: 377–395.
35. Tutz G, Gertheiss J. Feature extraction in signal regression: A boosting technique for functional data regression. Journal of Computational and Graphical Statistics 2010; 19: 154–174.
36. Slawski M, zu Castell W, Tutz G. Feature selection guided by structural information. The Annals of Applied Statistics 2010; 4: 1056–1080.
37. Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 2007; 35: 2313–2351.
38. Faschingbauer F, Beckmann M, Goecke T, Yazdi B, Siemer J, Schmid M, Mayr A, Schild RL. A new formula for optimized weight estimation in extreme fetal macrosomia (≥ 4500 g). European Journal of Ultrasound 2012; 33: 480–488.
39. Fenske N, Burns J, Hothorn T, Rehfuess EA. Understanding child stunting in india: A comprehensive analysis of socio-economic, nutritional and environmental determinants using additive quantile regression. PloS ONE 2013; 8: e78692.
40. Reiser V, Porzelius C, Stampf S, Schumacher M, Binder H. Can matching improve the performance of boosting for identifying important genes in observational studies? Computational Statistics 2013; 28: 37–49.
41. Rücker G, Reiser V, Motschall E, Binder H, Meerpohl JJ, Antes G, Schumacher M. Boosting qualifies capture-recapture methods for estimating the comprehensiveness of literature searches for systematic reviews. Journal of Clinical Epidemiology 2011; 64: 1364–1372.
42. Boosting (machine learning). From Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Boosting_(machine_learning); accessed July 7, 2014

43. Kearns M. Thoughts on hypothesis boosting. (http://www.cis.upenn.edu/~mkearns/papers/boostnote.pdf), unpublished manuscript (Machine Learning class project, December 1988).
44. Sariyar M, Borg A, Pommerening K. Evaluation of record linkage methods for iterative insertions. Meth Inf Med 2009; 48: 429– 437.
45. Stollhoff R, et al. An experimental evaluation of boosting methods for classification. Methods Inf Med 2010; 49: 219 – 229.
46. Sauerbrei W, Madjar H, Prömpeler HJ. Use of logistic regression and a classification tree approaches for the development of diagnostic rules: Differentiation of benign and malignant breast tumors based on color Doppler flow signals. Methods Inf Med 1998; 37: 226 –234.
47. Liu KE, Lo C-L, Hu Y-H. Improvement of adequate use of warfarin for the elderly using decision tree-based approaches. Methods Inf Med 2014; 53: 47–53.
48. Schmid M, Gefeller O, Hothorn T. Boosting into a new terminological era. Meth Inf Med 2012; 51 (2): 150 –151.
49. Dickersin K, et al. Identifying relevant studies for systematic reviews. BMJ 1994; 309: 1286 –1291.
50. Binder H, Schumacher M. Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. BMC Bioinformatics 2008; 9: 10–19.
51. Hofner B, Hothorn T, Schmid M, Kneib T. A Framework for Unbiased Model Selection Based on Boosting. Journal of Computational and Graphical Statistics 2012; 20 (4): 956 – 971.
52. Kim Y, Kim J. Gradient Lasso for feature selection. Proceedings of the 21st International Conference on Machine Learning. 2004.
53. Liu J, Huang J, Ma S. Incorporating network structure in integrative analysis of cancer prognosis data. Genetic Epidemiology 2013; 37: 173 –183.
54. Huang Y, Huang J, Shia BC, Ma S. Identification of cancer genomic markers via integrative sparse boosting. Biostatistics 2012; 13: 509–522.
55. Fahrmeir L, Tutz G. Multivariate Statistical Modelling based on Generalized Linear Models. New York: Springer-Verlag; 2001.
56. Tutz G. Regression for Categorical Data. Cambridge University Press; 2012.
57. Zahid FM, Tutz G. Multinomial logit models with implicit variable selection. Advances in Data Analysis and Classification 2013; 7: 393 – 416.
58. Zahid FM, Tutz G. Proportional odds models with high-dimensional data structure. International Statistical Review 2013; 81: 388 – 406.
59. Groll A, Tutz G. Regularization for generalized additive mixed models by likelihood-based boosting. Methods Inf Med 2012; 51 (2): 168–177.
60. Tutz G, Gertheiss J. Rating scales as predictors – the old question of scale level and some answers. Psychometrika 2014; 79 (3): 357–376.
61. Wang Z. HingeBoost: ROC-based boost for classification and variable selection. The International Journal of Biostatistics 2011; 7 (1): 1– 30.
62. Wang Z. Multi-class HingeBoost. Method and application to the classification of cancer types using gene expression data. Methods Inf Med 2012; 51 (2): 162 –167.
63. Vapnik V. The Nature of Statistical Learning Theory. New York: Springer-Verlag; 1996.
64. Meinshausen N, et al. p-Values for high-dimensional regression. Journal of the American Statistical Association 2009; 104 (488): 1671–1681.
65. Wang C, Feng Z. Boosting with missing predictors. Biostatistics 2010; 11 (2): 195 –212.
66. Tukey JD. Exploratory Data Analysis. Reading (MA): Addison-Wesley; 1977.
67. Mease D, Wyner A. Evidence contrary to the statistical view of boosting. J Mach Learn Res 2008; 9: 131–156.
68. König IR, et al. Predicting long-term outcome after acute ischemic stroke – a simple index works in patients from controlled clinical trials. Stroke 2008; 39: 1821– 1826.
69. Breiman L. Random Forests. Mach Learn 2001; 45: 5–32.
70. König IR, Malley JD, Weimar C, Diener H-C, Ziegler A, on behalf of the German Stroke Study Collaborators. Practical experiences on the necessity of external validation. Stat Med 2007; 26: 5499 –5511.

Addresses of the Authors

Peter Bühlmann
ETH Zürich
Seminar for Statistics
Rämistrasse 101, HG G 17
8092 Zürich
Switzerland
E-mail: buhlmann@stat.math.ethz.ch

Jan Gertheiss
Georg-August-University of Göttingen
Department of Animal Sciences
Biometrics & Bioinformatics Group
Carl-Sprengel-Weg 1
37075 Göttingen
Germany
E-mail: jgerthe@uni-goettingen.de

Stefanie Hieke
University Medical Center Freiburg
Department of Medical Biometry and Statistics
Center for Medical Biometry and Medical Informatics
Eckerstr. 1, room 107
79104 Freiburg
Germany
E-mail: hieke@imbi.uni-freiburg.de

Thomas Kneib
Georg-August-University of Göttingen
Chair of Statistics
Platz der Göttinger Sieben 5
37073 Göttingen, Germany
E-mail: tkneib@uni-goettingen.de

Shuangge Ma
Yale School of Public Health
Department of Biostatistics
60 College ST, LEPH 206
New Haven, CT 06520
USA
E-mail: shuangge.ma@yale.edu

Martin Schumacher
University Medical Center Freiburg
Department of Medical Biometry and Statistics
Center for Medical Biometry and Medical Informatics
Stefan-Meier-Str. 26, room 01- 019
79104 Freiburg, Germany
E-mail: ms@imbi.uni-freiburg.de

Gerhard Tutz
Ludwig Maximilians University
Department of Statistics
Seminar of Applied Stochastics
Akademiestraße 1, room 457
80799 München
Germany
E-mail:
gerhard.tutz@stat.uni-muenchen.de

Ching-Yun Wang
Fred Hutchinson Cancer Research Center
Public Health Sciences Division
1100 Fairview Avenue N.
P.O. Box 19024
M2-B500
Seattle, Washington 98109-1024
USA
E-mail: cywang@fhcrc.org

Zhu Wang
Connecticut Children's Medical Center
Department of Research
282 Washington Street
Hartford, CT 06106
USA
E-mail: zwang@connecticutchildrens.org

Andreas Ziegler
University of Lübeck
University Medical Center Schleswig-Holstein, Campus Lübeck
Institute of Medical Biometry and Statistics
Ratzeburger Allee 160
23562 Lübeck, Germany
E-mail: ziegler@imbs.uni-luebeck.de