Rejoinder

Andreas Ziegler*,[1,2]

[1] Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck, Universitätsklinikum Schleswig-Holstein, Campus Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
[2] Zentrum für Klinische Studien, Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

Received 13 January 2014; revised 1 March 2014; accepted 2 March 2014

This is the reply to the discussion of the two companion articles on the theory and application of "probability estimation with machine learning methods for dichotomous and multicategory outcome" by Kruppa et al. (2014; 534–563 and 564–583). The five discussion papers are Binder (2014; 584–587), Boulesteix and Schmid (2014; 588–593), Shin and Wu (2014; 594–596), Simon (2014; 597–600), and Steyerberg et al. (2014; 601–606).

Keywords: Logistic regression; Machine learning; Penalization; Probability estimation.

The remarks received by all of the discussants demonstrate that there is no silver bullet for analyzing prediction models. I found their comments and critiques both challenging and stimulating, and I am extremely grateful to them all for their efforts and for submitting their commentaries that provide further valuable insights into probability estimation. Obviously, the comments address further aspects and provide opinions from different angles, influenced by personal experiences. Unfortunately, this rejoinder has to be left uncommented, last but not least for reasons of space. The comments and this rejoinder are aimed to provide the basis for controversial discussion that should be continued in the near future. Although I liked to reply to each of the comments, I decided for a summarizing reply focusing on the most important central and recurring themes of (1) interpretability, (2) software, computational transportability, and reproducible research, (3) probability estimation in $n \gg p$ problems versus $p \gg n$ problems, (4) choice of tuning parameters, and, finally, (5) the role of logistic regression.

1 Interpretability

Statistics traditionally deals with inference, including effect size estimation and statistical significance testing, while machine learning methods are generally used for classification, probability estimation, or predictive modeling (Simon, 2014). In machine learning, the interpretability of the model is generally considered less important; for many machines it is even impossible. It might be this lack of interpretability that led Steyerberg et al. (2014) to their statement that "logreg will remain the default modeling approach to probability estimation in medical risk prediction, especially when applied with modern approaches" and that machine learning techniques "may have a supplementary role, in highly complex problems and to provide a comparison to regression results". As pointed out by Boulesteix and Schmid (2014), it might be the perspective that "should probably be asked and answered much more clearly by statisticians and biomedical scientists in practice". In my view, this could be phrased differently: If the primary aim is to investigate the role of a specific independent variable, that is feature,

*Corresponding author: e-mail: ziegler@imbs.uni-luebeck.de, Phone: +49-451-500-2780, Fax: +49-451-500-2999

or to better understand the underlying biology, a classical parametric regression model is probably the model of choice. However, if the primary aim is the construction of a prediction model aiming at low error estimates, interpretability of the model is of a lower-level concern. A specific machine learning approach might be even preferable in this situation over a classical parametric regression method when its performance fits the purpose of estimation.

Regarding the problem of marker identification, Binder (2014) stated that "the main danger then is to miss or misjudge an important predictor of clinical events that could be useful for arriving at medical decisions". However, if the primary aim is to derive a model with excellent prediction performance, the variables included in the final model are less important than the overall model performance. Nevertheless, "the established markers might need to be given a special role in the training process". Binder (2014) and Steyerberg et al. (2014) criticized that uniform treatment of all potential predictors provided seems to be a feature of many machine learning approaches. As pointed out by Binder, the unequal treatment of independent variables is easily implemented for regression modeling approaches by considering the effect of the new markers adjusted for established predictors. I admit that variables cannot be selected preferentially in some of the machine learning methods. However, for random forests it has been possible to select features with varying probabilities as split variables (Biau, 2012). This is implemented in our own software Random Jungle (Kruppa et al., 2014b). Similarly, features can be weighted differently in nearest neighbor methods in the distance calculations. I agree that the use of external information is meaningful for low dimensional prediction problems, and modeling should follow the general principles discussed in detail, for example by Harrell (2001). However, high dimensional data problems might ask for different modeling strategies.

The two different perspectives interpretative and predictive modeling have also been discussed by Shmueli (2010). The more general perspective seems to be that the pure predictive probability estimation approach without additional interpretation is generally considered to be not important. However, as pointed out by Steyerberg et al. (2014), "we can usually specify various models, which all reasonably describe the data". One such well-known example from the literature is the study of Golub et al. (1999) who aimed at distinguishing between acute myeloid leukemia and acute lymphoblastic leukemia using gene expression data shortly after the availability of microarrays. Their prediction models varied between 10 and 200 independent variables, and all models were found to be equally accurate. As analytically shown by Hand (2006) using a simple and intuitive model, only minor improvements in predictions can generally be expected when variables are added to a prediction model.

Fifteen years after the pioneering Golub et al. paper, I would like to emphasize that the pure predictive approach has been used and is currently used also with a strong focus on commercial use. Examples, established in clinical practice (Scharl et al., 2012) are the diagnostic tests Oncotype DX® and MammaPrint® that are both based on complex gene expression signatures. Alternatives to these test systems include tests based on immunohistochemistry, and among the available products are IHC4 and Mammostrat®. Oncotype DX and MammaPrint are based on 21 and 70 genes, respectively, and utilize complex gene expression profiles that obviously lack interpretability. Both test systems clearly are in vitro diagnostic multivariate index assays and cleared by the Food and Drug Administration (FDA) for commercial use. In contrast, the protein expression measures in the so-called IHC4 test are based on ER/PR, HER2, and Ki-67, which have been extensively studied in the literature so that a test system including these markers is biologically meaningful. Irrespective of interpretability, the tests are used in applications because test results affect treatment decision.

The pure predictive approach has been taken before. One example is the so-called triple test that quantifies the risk of giving birth to an infant with Down's syndrome using quantitative levels of α-fetoprotein (AFP), human chorionic gonadotrophin (hCG), and unconjugated oestriol (uE_3) in the serum of pregnant women (Wald et al., 1988; Norgaard-Pedersen et al., 1990). In analogy to the immunohistochemistry approaches, the number of molecular biomarkers used in these tests is limited, and biological interpretations are available. However, the triple test that was used commercially until approximately a decade ago, was a black box because formulae for calculations were not made available. By using an extensive search of the entire space of AFP, hCG, and uE_3 values, we reconstructed the

formulae of the risk calculations (Viethen and Ziegler, 1998). Furthermore, we were able to demonstrate substantial differences in the estimated risks between the different risk assessment systems and a clear lack of model validation.

Of course, if a model is interpretable, it is more likely to be accepted by clinicians compared to any black box systems. However, concerning Oncotype DX and Mammaprint, model performance and clinical utility turn out to be the more important aspects.

The acceptance of black box systems, such as the classical triple test or the more recent multimarker tests Oncotype DX and Mammaprint might generally be higher for tests utilizing molecular information, such as DNA, epigenetic, gene expression, metabolomic, or general protein expression markers when compared with clinical variables. With molecular markers the user might be willing to accept the general interpretation of a genetic or molecular involvement without understanding all specific details. Thus, these tests are not entirely black but rather gray boxes with regard to interpretability. In any case, the predictive models need to be externally or temporally validated (Altman and Royston, 2000).

2 Software and computational transportability

If a model is a black box, it is important that corresponding software objects are available (Boulesteix and Schmid, 2014). The software should also be simple to use, and it needs to be validated. As a result, the FDA has released a guideline on general principles of software validation. Ideally, the software works independently of the users local operating system; otherwise, the entire system needs to be validated to prevent errors such as the classical Microsoft Windows Calculator error (http://support.microsoft.com/kb/72540/en-us). If the model is not too complex, nomograms are an alternative to ready-to-use software packages. They are easy to use and provide in most cases the basis for a simple model interpretation. However, they involve some manual calculations, and such they are prone to human errors. Nevertheless, nomograms avoid the problems inherent to software validation, and allow an ease transportable.

Software validity also plays a role for the implementation of machine learning methods. Traditional regression models are available in many standard software packages (Boulesteix and Schmid, 2014), which enable a relatively simple validation. In contrast, implementations of sophisticated and fast machine learning approaches are not widespread, and the documentation could be rather cryptic. As some of the packages are in a development stage, upward compatibility is not always guaranteed. Furthermore, implementations may substantially differ between packages even within the same computing environment, and they may lead to inconsistent results. For example, differences have been identified in the implementation of importance measures for random forests (Schwarz et al., 2010). All these aspects make software validation of these learning machines difficult.

Computational transportability covers different aspects. As pointed out before (Kruppa et al., 2014a), the ability for model transportability differs substantially between machines. Specifically, nearest neighbor methods require the transfer of the entire training dataset (Boulesteix and Schmid, 2014; Kruppa et al., 2014a). The transfer is even more demanding for bootstrapped nearest neighbors. However, in some areas of research, such as gene expression analysis, the training data are often made publicly available anyway. Even more, several journals, including the *Biometrical Journal*, emphasize the reproducibility of research by asking authors to provide access to both data and computer code (Hothorn et al., 2009). This approach allows reproducing the calculations produced by others (Diggle and Zeger, 2010; Keiding, 2010). It also provides researchers with code that can be adapted for the analyses of other datasets although this might be a nontrivial task. In statistical methodological research, available code allows to understand all steps involved in a specific analysis. However, the availability of data and code does not allow reproducing the entire research because many decisions need to be made by the data monitoring and data management teams as well as the data analyst

before the master file that is used for analyses is available (Keiding, 2010; Peng, 2011; Anonymous, 2014).

When comparing random forests with nearest neighbors, transportability of a model is simpler for random forests because only the structure of all trees needs to be made available, and these can be easily displayed as a 2D array. The simplest representation is achieved for the standard logistic regression model, where only regression coefficients need to be made available. Boulesteix and Schmid (2014) note that the clear definition of all clinical variables and their measurement is important to transport the model. Unfortunately, this information also includes all data quality control and data transformation steps prior to the use of machine learning. To make these data available can be extremely challenging or even impossible for high throughput molecular data, when "low-level statistical analysis", such as transformations, filtering, and normalization procedures are applied in quality control steps prior to "high-level statistical analysis" (Ziegler et al., 2008). To give an example, quantile normalization is specific to a single training dataset. These obstacles hinder the transportability of a model, even when a standard regression model, such as logistic regression, is used.

3 Probability estimation in $n \gg p$ versus $p \gg n$ problems

Shin and Wu (2014) point out that most statistical methods including the probability estimation methods are not directly applicable in $p \gg n$ problems, that is when the number of features is large. Two different aspects need to be considered here.

First, more efficient implementations of machine learning methods for high dimensional data are urgently required. For example, some researchers have used support vector machines with linear kernels for $p \gg n$ problems, most likely because of the computational burden in case of the use of nonlinear kernels; see Shin and Wu (2014). In random forests, the computational complexity also increases substantially with the number of features available as split variable (`mtry`), and the choice of a high `mtry` value or even the tuning of `mtry` is difficult in most packages because of computational restraints in most implementations (Schwarz et al., 2010).

Second, $p \gg n$ problems seem to require different modeling strategies than $n \gg p$ problems, which have been studied over many years (Harrell et al., 1996; Harrell, 2001). Specifically, $n \gg p$ problems generally allow the explicit modeling of continuous independent variables, including restricted splines or fractional polynomials for dealing with nonlinearity (Steyerberg et al., 2014). Furthermore, independent variables may be selected carefully through a systematic literature search, and they may even be weighted according to external knowledge. This explicit modeling approach seems to be impractical in $p \gg n$ problems and is likely to yield unstable models. For example, a modern gene expression array, such as Affymetrix' Human Transcriptome Array 2.0 contains almost 300,000 transcripts, that is continuous independent variables. This would lead to more than 1 million parameters to be investigated when either cubic splines with five knots or fractional polynomials of degree 2 are used. Procedures for the automated stable identification of a reasonable nonlinear functional relationship between the feature and the outcome should avoid overfitting and should also be robust against outliers. The reliable identification of all variables predicting the endpoint in a multivariable model is, however, considered to be impossible because all samples have to be considered small according to the curse of dimensionality (Kruppa et al., 2014a).

As a result, it is important to distinguish between the identification of (1) a relevant variable for the inclusion in a model, (2) its functional form in a model, and (3) its causal influence. Phrased differently, whether an independent variable should be included in a model or not can be answered more stable than the question about its functional form. Whether the independent variable is causal is another question and related to causality; see for example the Hill criteria (Hill, 1965).

Feature selection has been identified in three commentaries as a major issue in $p \gg n$ problems (Binder, 2014; Shin and Wu, 2014; Steyerberg et al., 2014). For example, cost can be reduced substantially for molecular measurements in clinical routine if an array can be used which has been tailored to

the specific problem instead to a more global question that is mostly intended for research purposes with no commercial interest. Furthermore, analytic validity is generally higher for smaller microarrays and specific problems.

4 Tuning parameters

Steyerberg et al. (2014) states that machine learning techniques "are more flexible than regression, which makes them more data hungry. A technique such as [nearest neighbors] ... may be extreme in data requirements, because of its fully nonparametric nature". Díaz-Uriarte and Alvarez de Andrés (2006) proposed a backward feature selection method for random forests, and they suggested that there is little need to fine-tune the other parameters for excellent performance. Khondoker et al. (2013) and Kruppa et al. (2013) considered, however, various tuning parameters for random forests, such as the terminal nodesize (`nodesize`), the number of trees to be grown in a forest (`ntree`) in addition to `mtry`. It is intuitively clear that more trees are required in applications for probability estimation if the terminal nodesize is small. However, there are many more tuning options inherent to random forests. Specifically, it is unclear which split criterion should be used, whether bootstrap samples should be drawn with replacement or without replacement, how large the dataset from the resampling should be, or whether bootstrapping should be done with stratification by the proportion of samples in each category of the outcome variable (Ziegler and König, 2014). Furthermore, there are several conceptually different principles for generating random forests (Ziegler and König, 2014).

Parametric regression models, even more penalized regression models also allow for great flexibilities. First, one may choose between several regularization approaches, such as LASSO or elastic net. Second, in case of continuous variables, the functional form of the variable can be modeled in many different ways to allow for nonlinearity. Similarly, multicategory independent variables may be treated as continuous variables or several categories may be grouped. Third, interactions between variables can be integrated into the model as well and, finally, variable selection can also be performed. Penalized regression models might therefore also be "data hungry". Therefore, I agree more with Boulesteix and Schmid (2014) who stated that "the choice of the learning method, parameter tuning, and computational transportability are some of the remaining challenges that machine learning methods will have to address to establish themselves as prediction tools in biometrical practice". This applies, however, to both machine learning and classical regression methods.

5 Role of logistic regression

Steyerberg et al. (2014) grants machine learning approaches only a supplementary role for probability estimation in medical risk prediction. I fully agree with their interpretation for $n \gg p$ problems, where standard modeling approaches are available. However, in high dimensional data analysis machine learning methods may have more than a supplementary role although they clearly are computationally intensive (Shin and Wu, 2014).

Parametric models, such as logistic regression, have a higher rate of convergence and a lower variability, if the model is correctly specified. However, one should note that omission of variables leads to model misspecification in both the machine learning methods and parametric regression models. Machine learning methods are generally nonparametric and a misspecification of the functional form affects parametric regression models only. If the model is misspecified, the resulting maximum likelihood estimator still reaches a minimum and may be called minimum ignorance estimator (White, 1982). The relevance of the misspecification will, however, be application specific, and may be open for discussion.

An important advantage of parametric regression, such as logistic regression, over almost all machine learning approaches is, however, its calibration ability. Specifically, if the model is transferred to a new

setting with a different baseline probability, the original estimate of the intercept does not match the intercept of the new dataset anymore. With logistic regression, it is simple to only estimate the intercept on the new dataset, while keeping all other regression coefficients fixed. This is a simple way to calibrate the model (König et al., 2008). In machine learning, calibration generally is impossible, and approaches for model calibration are urgently required.

Acknowledgments A.Z. acknowledges funding from the European Union (BiomarCare, grant number: HEALTH-2011-278913), and the German Ministry of Education and Research (CARDomics, grant numbers: 01KU0908A and 01KU0908B; Phenomics, grant number: 0315536F). A.Z. gratefully thanks the four reviewers of the twin papers, the discussants, and Lutz Edler, editor of the *Biometrical Journal*. Their comments and guidance have greatly improved the presentation in the twin papers and this rejoinder.

Conflict of interest

A.Z. is statistical consultant for the Protagen AG, Dortmund. He is Associate Editor of the Biometrical Journal, Statistics in Medicine, and Methods of Information in Medicine.

References

Altman, D. G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine* **19**, 453–473.

Anonymous (2014). Number crunch. *Nature* **506**, 131–132.

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research* **13**, 1063–1095.

Binder, H. (2014). Discussion—what subject matter questions motivate the use of machine learning approaches compared to statistical models for probability prediction? *Biometrical Journal* **56**, 584–587.

Boulesteix, A. L. and Schmid, M. (2014). Discussion—machine learning versus statistical modeling. *Biometrical Journal* **56**, 588–593.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3.

Diggle, P. J. and Zeger, S. L. (2010). Embracing the concept of reproducible research. *Biostatistics* **11**, 375.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Science* **21**, 1–14.

Harrell, F. E. (2001). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York, NY.

Harrell, F. E., Jr., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361–387.

Hill, A. B. (1965). The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**, 295–300.

Hothorn, T., Held, L. and Friede, T. (2009). Biometrical journal and reproducible research. *Biometrical Journal* **51**, 553–555.

Keiding, N. (2010). Reproducible research and the substantive context. *Biostatistics* **11**, 376–378.

Khondoker, M., Dobson, R., Skirrow, C., Simmons, A. and Stahl, D. (2013). A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Statistical Methods in Medical Research*, doi: 10.1177/0962280213502437.

König, I. R., Ziegler, A., Bluhmki, E., Hacke, W., Bath, P. M. W., Sacco, R. L., Diener, H.-C., Weimar, C. and on behalf of the VISTA investigators (2008). Predicting long-term outcome after acute ischemic stroke—a simple index works in patients from controlled clinical trials. *Stroke* **39**, 1821–1826.

Kruppa, J., Liu, Y., Biau, G., Kohler, M., König, I. R., Malley, J. D. and Ziegler, A. (2014a). Probability estimation with machine learning methods for dichotomous and multi-category outcome: theory. *Biometrical Journal* **56**, 534–563.

Kruppa, J., Liu, Y., Diener, H. C., Holste, T., Weimar, C., König, I. R. and Ziegler, A. (2014b). Probability estimation with machine learning methods for dichotomous and multi-category outcome: Applications. *Biometrical Journal* **56**, 564–583.

Kruppa, J., Schwarz, A., Arminger, G. and Ziegler, A. (2013). Consumer credit risk: individual probability estimates using machine learning. *Expert Systems with Applications* **40**, 5125–5131.

Norgaard-Pedersen, B., Larsen, S. O., Arends, J., Svenstrup, B. and Tabor, A. (1990). Maternal serum markers in screening for Down syndrome. *Clinical Genetics* **37**, 35–43.

Peng, R. D. (2011). Reproducible research in computational science. *Science* **334**, 1226–1227.

Scharl, A., Thomssen, C. and Harbeck, N. (2012). AGO recommendations for diagnosis and treatment of patients with early and metastatic breast cancer: update 2012. *Breast Care* **7**, 322–335.

Schwarz, D. F., König, I. R. and Ziegler, A. (2010). On safari to random jungle: a fast implementation of random forests for high dimensional data. *Bioinformatics* **26**, 1752–1758.

Shin, S. J. and Wu, Y. (2014). Discussion—variable selection in large margin classifier-based probability estimation with high-dimensional predictors. *Biometrical Journal* **56**, 594–596.

Shmueli, G. (2010). To explain or to predict? *Statistical Science* **25**, 289–310.

Simon, R. (2014). Discussion—class probability estimation for medical studies. *Biometrical Journal* **56**, 597–600.

Steyerberg, E. W., van der Ploeg, T. and Van Calster, B. (2014). Discussion—risk prediction with machine learning and regression methods. *Biometrical Journal* **56**, 601–606.

Viethen, G. and Ziegler, A. (1998). Application of triple test in prenatal diagnostics—an evidence-based approach. In: Bremen Greiser, E. and Wischnewsky, M. (Eds.), *GMDS '98*, CD 1–20, Medien und Medizin-Verlag, Munich, Germany.

Wald, N. J., Cuckle, H. S., Densem, J. W., Nanchahal, K., Royston, P., Chard, T., Haddow, J. E., Knight, G. J., Palomaki, G. E. and Canick, J. A. (1988). Maternal serum screening for Down's syndrome in early pregnancy. *BMJ* **297**, 883–887.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Ziegler, A. and König, I. R. (2014). Mining data with random forests: current options for real-world applications. *Wiley Interdisciplinary Reviews* **4**, 55–63.

Ziegler, A., Thompson, J. R. and König, I. R. (2008). Biostatistical aspects of genome-wide association studies. *Biometrical Journal* **50**, 8–28.

