Discussion

Risk prediction with machine learning and regression methods

Ewout W. Steyerberg[*,1], **Tjeerd van der Ploeg**[2], and **Ben Van Calster**[3]

[1] Department of Public Health, Erasmus MC, Rotterdam, The Netherlands
[2] Medical Centre Alkmaar/Inholland University, Alkmaar, The Netherlands
[3] Department of Development and Regeneration, KU Leuven, Leuven, Belgium

Received 20 December 2013; revised 10 January 2014; accepted 10 January 2014

This is a discussion of issues in risk prediction based on the following papers: "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory" by Jochen Kruppa, Yufeng Liu, Gérard Biau, Michael Kohler, Inke R. König, James D. Malley, and Andreas Ziegler; and "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Applications" by Jochen Kruppa, Yufeng Liu, Hans-Christian Diener, Theresa Holste, Christian Weimar, Inke R. König, and Andreas Ziegler.

Keywords: Machine learning; Prediction; Regression.

Cross-fertilization between medical statistics and epidemiology on the one hand and machine learning techniques (MLT) on the other can be very stimulating (Kruppa et al., 2014a; Kruppa et al., 2014b). Probability estimation is key to the area of risk prediction, which is growing in importance in medicine, where personalized medicine becomes more and more possible through the combination of classical risk predictors and biomarkers.

The first paper focuses on theoretical aspects, such as consistency of probability estimation (Kruppa et al., 2014a). For example, for the nearest neighbor (NN) method the authors report that the error in the estimation of probabilities converges to zero if certain assumptions are met and the sample size tends to infinity, while this is not strictly true for random forests (RF). Consistency does not hold for logistic regression (logreg), where the validity of probability estimates depends on the model specification. Simulation studies are provided which show that each of these methods can fail to provide reasonable predictions. Calibration properties were particularly poor for some variants of support vector machines (SVMs) in some simulations, i.e. poor agreement between true probability and predicted probability. Various performance criteria were studied, specifically squared scoring rules such as the Brier score. Rank-based measures such as the area under the ROC curve were also used by Kruppa et al., for which extensions to multicategory evaluation have recently been proposed, such as the Polytomous Discrimination Index (Van Calster et al., 2012). Likelihood based performance measures might also have been be used, such as Nagelkerke's R^2 (or other variants, Austin and Steyerberg, 2013), but these would probably have led to the same impression of performance. Finally, the paper illustrates that some methods behave very similarly, e.g. two variants of NN, and SVM with linear kernel and logreg.

Below we first discuss tuning and implementation aspects of machine learning techniques (MLT) and regression models (Section 1), followed by reflections on model uncertainty (Section 2) and a

*Corresponding author: e-mail: e.steyerberg@erasmusmc.nl, Phone: +31-10-7030045

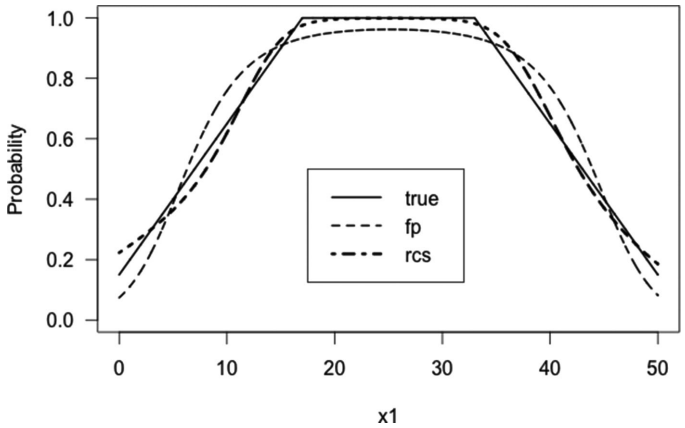

Figure 1 The probability of $y = 1$ in Simulation study I, for $x_2 = 25$. We simulated 5000 subjects, where the selected FP function was $(x_1+0.1)+(x_1+0.1)^2$. The rcs function used 5 knots (4 *df*).

possibly sensible modeling strategy (Section 3). We end with reflections on the potential role of MLT in addition to regression modeling (Section 4).

1 Tuning, traditions, and modern approaches in regression modeling

One issue of attention with MLT is that they have various tuning parameters. These include the number of neighbors to consider in NN, the regularization parameters and type of kernel for SVM, and the tree specifications for RF, which essentially serve to control the complexity of the fitted model. Similarly, various strategies and modeling approaches are possible for logreg.

First, prediction modelers of medical data should assess nonlinearity of continuous variables (Harrell, 2001). The blind application of the logistic regression model $y \sim x_1+x_2$, as was presented in Simulation I, is not very realistic. The underlying circle model requires some kind of increasing and decreasing functions for x_1 and x_2. Any epidemiologist would do some form of data inspection, and would immediately note the more or less squared relation with x_i (Fig. 1). Preferences for modeling non-linearity vary: Harrell advices restricted cubic splines (rcs) as a default tool in regression modeling (2001), while Royston and Sauerbrei (2008) advocate the use of fractional polynomials (FP). For illustration, we fitted FP and rcs functions in a simulation with 5000 subjects (Fig. 1, using R packages mfp and rms). The true effect of x_1 is a linear increase from $x_1 = 0$ to $x_1 = 17$, a probability of 1 between $x_1 = 17$ and $x_1 = 33$, and a linear decrease between $x_1 = 33$ and $x_1 = 50$. For the FP model, a linear term plus square term are selected for x_1. This FP model follows the true shape well, although the probability of 1 is not reached, and low probabilities are underestimated. The rcs model (with 5 knots, 4 *df*) reached the plateau probability of 1, but slightly overestimated low probabilities at $x_1 = 0$ and $x_1 = 50$. The models $y \sim fp(x_1) + fp(x_2)$ and $y \sim rcs(x_1) + rcs(x_2)$ had Brier scores below 0.15, which is equivalent to the best performing MLTs in this simulation (NN, SVM-Bessel). So, as may be expected, a reasonably specified logreg model performs very well in simulation I.

Second, whereas some form of regularization is indispensable for MLT due to their flexibility, similar techniques exist for logreg to penalize or shrink model coefficients. Examples are L1 (LASSO) or L2 (ridge) penalization, or Bayesian approaches. The LASSO method uses a L1 penalty to shrink regression coefficients to zero (Tibshirani, 1996). Hence LASSO combines variable selection with shrinkage while still providing adequate predictions, as observed in a large simulation study for patients with an acute myocardial infarction (Steyerberg et al., 2000). Similar to the improvement of

RF over CART for prediction (Austin et al., 2012), we should use penalized rather than traditional approaches for logreg if comparisons are made between logreg and MLT.

2 Model uncertainty and parsimony

A major problem for prediction models is model uncertainty. We can usually specify various models, which all reasonably describe the data (Breiman, 2001). In medical research, we may often have a relatively long list of potential predictors, e.g. 49 for Application I (stroke) (Kruppa et al., 2014b). This list was apparently based on solid grounds (a systematic literature review), but some reduction might have been possible by posing stricter criteria on the evidence underpinning a potential prognostic effect, such as consistency of a substantial effect size across multiple studies. It is not plausible that a medical problem has 49 equally important predictors (where we recognize that ''importance'' may depend on the modeling technique used). For example, we identified only 3 key predictors of 6 months outcome in a systematic literature review for patients with traumatic brain injury (Mushkudiani et al., 2008). In this prediction problem, Age, Glasgow Coma Scale—especially the Motor component—and pupillary reactivity strongly predicted 6 months mortality (Steyerberg et al., 2008). Models with these key predictors performed well in temporal and geographical validations (Roozenbeek et al., 2012). Only minor improvements were noted by including other characteristics, such as CT scan findings, while many clinicians would consider these characteristics vital for prediction.

Moreover, it is well known that medical data typically have a poor ''signal to noise ratio'' for predictors. This has two implications. First, sample size and penalization are key factors to accurate prediction modeling. This is true for regression models, and even more so for MLT. MLT are more flexible than regression, which makes them more data hungry. A technique such as NN may be extreme in data requirements, because of its fully nonparametric nature. Second, more parsimonious model specifications may often be sufficient to capture the main structure of a prediction problem. Extreme nonlinearity such as in the presented Simulation I is infrequently observed in medical research. Complex higher order interactions may occasionally exist but impossible to identify in reasonably sized medical data sets. This is supported by recent studies that report similar performance of logreg versus MLT (Van Calster et al., 2009, 2010; Van der Ploeg et al., 2011; Austin et al., 2013).

3 Sensible prediction modeling in medical data?

Medical data sets are often of too small size to be able to reliably address difficult research questions, such as determining which predictors are important and which are not. For example, reliably determining which of 49 characteristics predict mortality may require far larger numbers of events than occurring among the training set of 1737 patients in Application I (Kruppa et al., 2014b). In addition, backward elimination is a standard approach for variable selection in regression analysis, commonly assessed using $p < 0.05$ for predictors in a prediction model. Many drawbacks have been discussed in the past, including biased estimation of regression coefficients, distortion of the estimation of variance and p-values, and instability of the selected set of predictors (Austin and Tu, 2004; Sauerbrei and Schumacher, 1992; Steyerberg et al., 1999). For probability estimation the most relevant issue is that stepwise selection leads to suboptimal prediction: only the most prominent predictors are selected, so information from close-to-significant predictors is lost, and effects are exaggerated, which leads to too extreme predictions.

Sensible modeling should find a balance between external knowledge from outside the data versus what can be learned from the data. The smaller the data set available, the more we have to rely on external information. This holds primarily for the list of candidate predictors in a model, which is relevant to both MLT and logreg. But it also holds for issues such as whether we should rely on the additivity assumption in logreg, i.e. whether we should consider statistical interaction terms. Some

Table 1 Characteristics of MLT and regression modeling techniques.

Method	Consistency	Flexibility	Sample size requirements	Interpretability
NN	+	+	−	−
RF	+/−	+	+/−	−
SVM	+/−	+	+/−	−
Logreg	−	+/−	+	+

NN: nearest neighbors; RF: random forest using probability estimation trees; SVM: support vector machine; logreg: logistic regression.

traditional statisticians might consider assessment of interactions as good modeling practice, while others would warn for overfitting by the potential for inclusion of spurious interactions. Findings in prior studies and sample size of the data under study are key considerations for such strategies (Steyerberg, 2009).

4 A role for MLT in addition to regression?

MLT have various attractive properties such as its focus on regularization and on finding algorithms and classification models that work, rather than focusing strongly on theory of an assumed stochastic data model (Breiman, 2001). Clinical risk prediction research uses a similar philosophy, focusing on performance issues such as discrimination, calibration, utility, and impact. Nevertheless MLT also have various problems. If we aim for an important role of prediction models in medicine, we need to follow a framework that not only includes model development, but involves a process of validation and updating of models (Steyerberg et al., 2013). Updating may require adjustments to local settings (van Houwelingen, 2000). In logreg, simple updating to the average probability is easily achieved by changing the model intercept, while this is difficult for MLT.

Furthermore, interpretability to a clinical audience is usually essential (Kruppa et al., 2014b). Logistic regression models can transparently be presented, with insight in the relative effects of predictors by odds ratios and in nomograms, score charts and other displays. Such presentations are not possible for MLTs, although efforts to this end have been undertaken (Van Belle et al., 2012). We however notice that models are increasingly implemented on the internet. For example, a risk calculator for the probability of Lynch syndrome related mutation is accessed over 1000 times a month (Kastrinos et al., 2011). Web-based calculation of risk may allow the underlying model to be quite complex, e.g. a MLT.

Some characteristics of MLT and regression modeling techniques are summarized in Table 1. An NN approach may be attractive because of the theoretical property of consistency, but is data hungry (requires huge sample sizes) and lacks interpretability, similar to RF and SVM. The consistency of RF and SVM may not fully be proven, but the flexibility is large. Although logreg is not consistent in the estimation of probabilities, the flexibility can be substantial with a modern modeling strategy. Naïve fitting of linear main effects and automatic selection methods such as backward stepwise selection with $p < 0.05$ are suboptimal default implementations of logreg. Nonlinear transformations can readily be made by rcs and FP functions, and the shrinkage or penalization methods such as LASSO provide better than standard predictive performance. Sample size requirements for logreg depend on how much external evidence is available, and how much the analyst is willing to rely on such evidence, e.g. on the relevance and effects of predictors. Interpretability of effect sizes is readily possible by a medically trained audience, and model updating can readily be achieved with simple or more advanced procedures.

All in all, we envision that logreg will remain the default modeling approach to probability estimation in medical risk prediction, especially when applied with modern approaches. MLT may have a supplementary role, in highly complex problems and to provide a comparison to regression results.

Conflict of interest
The authors have declared no conflict of interest.

References

Austin, P. C. and Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology* **57**, 1138–1146.

Austin, P. C., Lee, D. S., Steyerberg, E. W. and Tu, J. V. (2012). Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods? *Biometrical Journal* **54**, 657–673.

Austin, P. C. and Steyerberg, E. W. (2013). Predictive accuracy of risk factors and markers: a simulation study of the effect of novel markers on different performance measures for logistic regression models. *Statistics in Medicine* **32**, 661–672.

Austin, P. C., Tu, J. V., Ho, J. E., Levy, D. and Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of Clinical Epidemiology* **66**, 398–407.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science* **16**, 199–215.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, New York, NY.

Kastrinos, F., Steyerberg, E. W., Mercado, R., Balmana, J., Holter, S., Gallinger, S., Siegmund, K. D., Church, J. M., Jenkins, M. A., Lindor, N. M., Thibodeau, S. N., Burbidge, L. A., Wenstrup, R. J. and Syngal, S. (2011). The PREMM(1,2,6) model predicts risk of MLH1, MSH2, and MSH6 germline mutations based on cancer history. *Gastroenterology* **140**, 73–81.

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 multicategory 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 multicategory outcome: Applications. *Biometrical Journal* **56**, 564–583.

Mushkudiani, N. A., Hukkelhoven, C. W., Hernandez, A. V., Murray, G. D., Choi, S. C., Maas, A. I. and Steyerberg, E. W. (2008). A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. *Journal of Clinical Epidemiology* **61**, 331–343.

Perel, P., Arango, M., Clayton, T., Edwards, P., Komolafe, E., Poccock, S., Roberts, I., Shakur, H., Steyerberg, E. and Yutthakasemsunt, S. (2008). Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *British Medical Journal* **336**, 425–429.

Roozenbeek, B., Lingsma, H. F., Lecky, F. E., Lu, J., Weir, J., Butcher, I., McHugh, G. S., Murray, G. D., Perel, P., Maas, A. I. and Steyerberg, E. W. (2012). Prediction of outcome after moderate and severe traumatic brain injury: external validation of the IMPACT and CRASH prognostic models. *Critical Care Medicine* **40**, 1609–1617.

Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-Building: a Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. John Wiley, Chichester, UK.

Sauerbrei, W. and Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* **11**, 2093–2109.

Steyerberg, E. W., Eijkemans, M. J. and Habbema, J. D. (1999). Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology* **52**, 935–942.

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., Jr. and Habbema, J. D. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine* **19**, 1059–1079.

Steyerberg, E. W., Mushkudiani, N., Perel, P., Butcher, I., Lu, J., McHugh, G. S., Murray, G. D., Marmarou, A., Roberts, I., Habbema, J. D. and Maas, A. I. (2008). Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Medicine* **5**, e165.

Steyerberg, E. W. (2009). *Clinical prediction models: a practical approach to development, validation, and updating*. Springer, New York, NY.

Steyerberg, E. W., Moons, K. G., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., Altman, D. G. and Group, P. (2013). Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS Medicine* **10**, e1001381.

Tibshirani, R. (1996). Regression and shrinkage via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

Van Belle, V. M., Van Calster, B., Timmerman, D., Bourne, T., Bottomley, C., Valentin, L., Neven, P., Van Huffel, S., Suykens, J. A. and Boyd, S. (2012). A mathematical model for interpretable clinical decision support with applications in gynecology. *PLoS One* **7**, e34312.

Van Calster, B., Condous, G., Kirk, E., Bourne, T., Timmerman, D. and Van Huffel, S. (2009). An application of methods for the probabilistic three-class classification of pregnancies of unknown location. *Artificial Intelligence in Medicine* **46**, 139–154.

Van Calster, B., Valentin, L., Van Holsbeke, C., Testa, A. C., Bourne, T., Van Huffel, S. and Timmerman, D. (2010). Polytomous diagnosis of ovarian tumors as benign, borderline, primary invasive or metastatic: development and validation of standard and kernel-based risk prediction models. *BMC Medical Research Methodology* **10**, 96.

Van Calster, B., Van Belle, V., Vergouwe, Y., Timmerman, D., Van Huffel, S. and Steyerberg, E. W. (2012). Extending the c-statistic to nominal polytomous outcomes: the Polytomous Discrimination Index. *Statistics in Medicine* **31**, 2610–2626.

Van der Ploeg, T., Smits, M., Dippel, D. W., Hunink, M. and Steyerberg, E. W. (2011). Prediction of intracranial findings on CT-scans by alternative modelling techniques. *BMC Medical Research Methodology* **11**, 143.

Van Houwelingen, H. C. (2000). Validation. calibration, revision and combination of prognostic survival models. *Statistics in Medicine* **19**, 3401–3415.

