

Statistics versus machine learning: definitions are interesting (but understanding, methodology, and reporting are more important)

Ben Van Calster^{1,2}, Jan Y Verbakel^{3,4}, Evangelia Christodoulou¹, Ewout W Steyerberg², Gary S Collins^{5,6}

1 KU Leuven, Department of Development and Regeneration, Leuven, Belgium; 2 Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, the Netherlands; 3 KU Leuven, Department of Public Health and Primary Care, Leuven, Belgium; 4 Nuffield Department of Primary Care Health Sciences, University of Oxford, UK; 5 Centre for Statistics in Medicine, Botnar Research Centre, University of Oxford, Oxford, UK; 6 NIHR Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK.

We thank Bian and colleagues for their interest in our study. Since the study was published, the distinction between logistic regression and machine learning has fueled a lot of discussion. We had addressed this issue already in the initial submission, but corroborated on it based on the reviewers' comments. We state in the paper that we do not believe there is a clear dichotomy, but rather that algorithms lie on a continuum regarding flexibility, and reliance on the data versus subject knowledge. Nevertheless, several publications and discussions explicitly make this distinction, and often conclude that machine learning leads to better predictive performance compared to traditional statistical methods. This justifies the pragmatic definition in our paper.

We feel that discussing semantics and definitions can be insightful, but it should not be the focal point resulting from our study. The key messages of our study on clinical risk prediction modeling are: (1) by itself, using highly flexible algorithms do not necessarily lead to improved performance, (2) methodological conduct in developing, validating and fair comparison of different algorithms needs to improve and (3) the reporting of prediction model studies should adhere to current guidelines such as TRIPOD.¹ The sensible development of a prediction model depends on the specific context, should bear in mind the clinical setting in which the algorithm is intended to be used, and needs to be carefully and fully described.

Nevertheless, we respect the comments on the practical choices that we made in our study. We compared 'logistic regression' with 'machine learning'. The category 'logistic regression' included standard maximum likelihood logistic regression, and penalized logistic regression (lasso, ridge, elastic net), but excluded bagged/boosted logistic regression and other algorithms that we labeled as traditional statistical methods. Penalization (regularization) has origins deep in the statistical literature. We refer to Stein's paradox, described in 1955 at the Berkeley Symposium on Mathematical Statistics and Probability.² This inspired the development of penalized linear regression methods such as ridge regression in 1970 and the lasso in 1996.^{3,4} The category 'machine learning' included everything except the 'logistic regression' category or other traditional statistical methods. We agree that support vector machines with linear kernels or Naïve Bayes have limited flexibility by design. Hence it makes sense to rank algorithms by complexity or flexibility.⁵ This is informative, and helps researchers to choose an algorithm that is reasonable for the specific purpose at hand, and in balance with the amount of data and subject knowledge available.

References

1. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD statement. *J Clin Epidemiol* 2015;68:112-21.
2. Stein C. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. In: J Neyman, editor. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, Berkeley: University of California Press; 1956, p. 197-206.
3. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970;12:55-67.
4. Tibshirani R. Regression Shrinkage and Selection via the Lasso. *J R Stat Soc B* 1996;58:267-88.

5. Ye J. On measuring and correcting the effects of data mining and model selection. *J Am Stat Assoc* 1998;93:120-31.