References

[1] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12−22.

[2] Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 2001;16:199−231.

[3] Mitchell TM. Machine learning. 1st ed. New York, NY: McGraw-Hill Education; 1997.

[4] Boulesteix A-L, Schmid M. Machine learning versus statistical modeling. Biom J 2014;56:588−93.

[5] Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317−8.

[6] Cramer JS. The early origins of the logit model. Stud Hist Philos Sci C 2004;35:613−26.

[7] Vapnik VN. The nature of statistical learning theory. New York, NY: Springer New York; 2000.

[8] Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: Dietterich TG, Becker S, Ghahramani Z, editors. Advances in neural information processing systems 14. Cambridge, MA: MIT Press; 2002:841−8.

[9] Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform 2002;35:352−9.

[10] Bergadano F, Cutello V, Gunetti D. Abduction in machine learning. In: Gabbay DM, Kruse R, editors. Abductive reasoning and learning. Dordrecht: Springer Netherlands; 2000:197−229.

[11] Prosperi MCF, Altmann A, Rosen-Zvi M, Aharoni E, Borgulya G, Bazso F, et al. Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. Antivir Ther 2009;14:433−42.

[12] Fraccaro P, Nicolo M, Bonetto M, Giacomini M, Weller P, Traverso CE, et al. Combining macula clinical signs and patient characteristics for age-related macular degeneration diagnosis: a machine learning approach. BMC Ophthalmol 2015;15:10.

[13] Vapnik VN, Chervonenkis AY. On the uniform convergence of relative frequencies of events to their probabilities. In: Vovk V, Papadopoulos H, Gammerman A, editors. Measures of complexity: festschrift for alexey chervonenkis. Cham: Springer International Publishing; 2015:11−30.

[14] Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak 2018;18:139.

[15] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206−15.

http://dx.doi.org/10.1016/j.jclinepi.2019.08.003

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

We thank Bian et al 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 article 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 with traditional statistical methods. This justifies the pragmatic definition in our article.

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 as follows: (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 transparent reporting of a multivariable prediction model for individual prognosis or diagnosis [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.

Ben Van Calster*
Department of Development and Regeneration
KU Leuven
Leuven, Belgium
Department of Biomedical Data Sciences
Leiden University Medical Centre
Leiden, The Netherlands

Jan Y. Verbakel
Department of Public Health and Primary Care
KU Leuven
Leuven, Belgium

DOI of original article: http://dx.doi.org/10.1016/j.jclinepi.2019.08.003.

Conflict of interest: none.

Funding: Research Foundation−Flanders (FWO) [grant G0B4716N]; Internal Funds KU Leuven [grant C24/15/037]; Cancer Research UK [grant 5529/A16895]; the NIHR Biomedical Research Centre, Oxford, UK.

Nuffield Department of Primary Care Health Sciences
University of Oxford
UK

Evangelia Christodoulou
Department of Development and Regeneration
KU Leuven
Leuven, Belgium

Ewout W. Steyerberg
Department of Biomedical Data Sciences
Leiden University Medical Centre
Leiden, The Netherlands

Gary S. Collins
Centre for Statistics in Medicine
Botnar Research Centre
University of Oxford
Oxford, UK
NIHR Oxford Biomedical Research Centre
John Radcliffe Hospital
Oxford, UK

*Corresponding author. Tel.: +32 16377788; fax: +32 16344205.
E-mail address: ben.vancalster@kuleuven.be (B. Van Calster)

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: Neyman J, editor. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. Berkeley, CA: University of California Press; 1956:197–206.

[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 Series B Stat Methodol 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.

http://dx.doi.org/10.1016/j.jclinepi.2019.08.002

Validity of health transition questions is supported by larger clinical improvements in purposive samples enriched for improvers

To the editors,

Health transition questions, which ask patients to report if they are better, worse, or unchanged after an intervention, are commonly used as anchors in determining clinically important changes in health outcome measures [1]. However, the validity of health transition questions has rarely been examined [2,3]. We previously showed that responses to health transition questions correlate with changes in self-rated health based on clinical vignettes [4]. As an additional test of the construct validity of health transition questions, we examined if responses on health transition questions corresponded with improvements in outcome measures. Specifically, we tested if samples enriched with higher proportions of patients who reported improvement on the transition question (as opposed to no change or worsening) demonstrated larger effect sizes than samples with lower proportions of patients who reported improvement.

What is new?

- Samples enriched to include more patients who reported improvement on a health transition question demonstrated larger effect sizes.
- These results support the validity of health transition questions.
- Transition questions may continue to be used as anchors for assessing clinically important improvement.

We used data from an observational study of treatment responses in 250 patients with active rheumatoid arthritis [5]. Patients were examined before and after treatment escalation for changes in several measures, including self-reported pain severity (by visual analog scale), physical function (by the SF-36 physical function subscale), swollen joint count (by physician), and the Simplified Disease Activity Index (SDAI), a composite measure of joint swelling, tenderness, patient global assessment, physician global assessment, and C-reactive protein level. After treatment, patients reported whether they had improved or not on domain-specific transition questions, including ones for pain, physical ability, joint swelling, and overall arthritis status [6]. The wording of the transition question was “Since the start of the study, my (pain/ability to do things/joint swelling/overall my arthritis) has: improved, stayed the same, or gotten worse.”

From the 250 patients, we drew 200 random samples of 100 patients each, first selected so that 20 patients reported improvement on the pain transition question and 80 patients reported worsening or no change. We used standardized response means (SRM = mean change/change standard deviation) as the effect size measure. We computed SRMs for the pain visual analog scale for each random sample and computed the mean of the 200 results. We then successively repeated this procedure for samples specified to have 25% of patients who reported improvement through samples specified to have 90% of patients who reported improvement. We then repeated this process for physical functioning, swollen joint count, and SDAI. SRMs for each measure increased progressively as the percent of patients who reported improvement on the

Declarations of interest: None.
Conflicts of interest: None.