

Validation stands on three feet, not two

Richard J Stevens (0000-0002-9258-4060) and Katrina K Poppe.

Nuffield Department of Primary Care Health Sciences, University of Oxford.

Richard Stevens, Associate Professor.

Faculty of Medical and Health Sciences, University of Auckland.

Katrina K Poppe, Senior Research Fellow,

Correspondence to: R Stevens, richard.stevens@phc.ox.ac.uk

We welcome Junfeng Wang's response to our article on the calibration slope, and for providing further evidence of misleading terms and concepts in this area. As Wang observes the slope of a calibration plot "actually has little relation with calibration", the discrimination slope suffers from a similar fate, and "new names are needed to prevent further misuse and misunderstanding of these measures" (1). Wang writes about models that have linear predictors, such as logistic and Cox regression, but the observations apply more generally.

Wang takes a view that two statistics are mismatched with two "aspects" of validation. We agree about the mismatch. Our understanding of the problem however is that there are really three aspects of validation, and that this is why definitions of the two terms calibration and discrimination can overlap.

Bias, ordering and spread are three concepts that do not overlap: it is possible to vary any one of these while the others remain constant (for example, adding a constant to the risk estimates will change the bias of a risk predictor, while leaving ordering and spread unchanged). We can then see that some texts have equated calibration with bias, while others have equated it with bias and spread (2,3). Similarly, while some authors including Wang equate discrimination strongly with ordering, others take discrimination to include both ordering and spread (2).

Bias and ordering are mathematical concepts for which mathematical definitions exist, whereas calibration and discrimination have historically been defined verbally (4-6). More recently, Van Calster et al. provide rigorous definitions of multiple levels of calibration: we note that the lowest level in this hierarchy is defined by bias, the second level is defined by bias and spread, and the third level requires low bias, correct ordering and good spread (7).

For the special case of generalised linear models, Wang argues that discrimination depends on linear predictors alone but calibration depends on the absolute probabilities. We would expand on this to reply that ordering depends on linear predictors alone, and bias and spread depend on the absolute probabilities. We also remark that the concepts of bias, ordering and spread, and Van Calster's hierarchy, extend beyond generalised linear models to any risk calculator, regardless of whether the algorithm is based on a model or on a non-parametric learning method.

(1) Wang J. Calibration slope versus discrimination slope: shoes on the wrong feet. J Clin Epidemiol 2020; in press.

(2) Harrell Jr. FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15(4):361-387.

(3) Vach W. Calibration of clinical prediction rules does not just assess bias. *J Clin Epidemiol* 2013;66(11):1296-1301.

(4) Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19(4):453-473.

(5) Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating. New York ; London: Springer; 2009.

(6) Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68(3):279-289.

(7) Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: From utopia to empirical data. *J Clin Epidemiol* 2016;74:167-176.