

ORIGINAL ARTICLE

Validation of clinical prediction models: what does the “calibration slope” really measure?

Richard J. Stevens^{a,*}, Katrina K. Poppe^b^a*Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK*^b*Faculty of Medical and Health Sciences, University of Auckland, Auckland, New Zealand*

Accepted 19 September 2019; Published online 9 October 2019

Abstract

Background and Objectives: Definitions of calibration, an aspect of model validation, have evolved over time. We examine use and interpretation of the statistic currently referred to as the calibration slope.

Methods: The history of the term “calibration slope”, and usage in papers published in 2016 and 2017, were reviewed. The behaviour of the slope in illustrative hypothetical examples and in two examples in the clinical literature was demonstrated.

Results: The paper in which the statistic was proposed described it as a measure of “spread” and did not use the term “calibration”. In illustrative examples, slope of 1 can be associated with good or bad calibration, and this holds true across different definitions of calibration. In data extracted from a previous study, the slope was correlated with discrimination, not overall calibration. Many authors of recent papers interpret the slope as a measure of calibration; a minority interpret it as a measure of discrimination or do not explicitly categorise it as either. Seventeen of thirty-three papers used the slope as the sole measure of calibration.

Conclusion: Misunderstanding about this statistic has led to many papers in which it is the sole measure of calibration, which should be discouraged. © 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Keywords: Clinical prediction rule; Calibration; Validation; Discrimination; Spread; Slope

1. Introduction

When clinical prediction models, including algorithms, regression models, and risk scores, are intended for diagnosis or prognosis, a key step on the pathway to clinical use is validation: demonstration that the model makes useful predictions in a target population, and not only the data on which it was derived [1,2]. The term “discrimination” is used for the ability of a model to distinguish people with

different true outcomes [3]. Another necessary aspect of validation is to evaluate “calibration”: the degree to which numerical predictions are too high or too low compared to outcomes [3]. Definitions of calibration vary and have evolved over time. Earlier papers define calibration to be “bias” in the statistical sense: the extent to which predictions, on average, agree with overall outcomes in the sample or population [4,5]. We will refer to this as “overall calibration.” More recent papers identify calibration as a property to be satisfied in increasingly fine subgroups or individuals [6–9]. In this approach, calibration is a multidimensional concept that cannot be summarized in a single statistic. Therefore, authors often recommend that two statistics known as the calibration intercept and calibration slope be used. If the intercept is close to 0 and the slope is close to 1, then there is good overall calibration and good calibration across a range of risk groups [4]. In practice, however, the statistic known as the calibration slope is sometimes reported in isolation.

We examine use and interpretation of the statistic currently referred to as the calibration slope.

The statistic known as the calibration slope does not by itself measure the calibration of a clinical prediction model, and the misleading term “calibration slope” should be dropped to promote more appropriate reporting of this and related statistics.

Conflicts of interest: The authors have no conflict of interest to declare. Funding: None.

Data sharing statement: The data in Table 1 are public domain data obtained by systematic review. The simulated data in Figure 1 were created by code that can be obtained from the corresponding author on request. The data in Figure 2 were extracted from Reference 16, Supplementary Material 3a, columns headed Strategy (2).

* Corresponding author. Tel.: +44 1865 289355; fax: +44 1865 270708.

E-mail address: richard.stevens@phc.ox.ac.uk (R.J. Stevens).

What is new?

Key findings

- Many authors equate slope of 1 to good calibration, omitting to calculate and report the intercept.
- If the intercept or equivalent information has not been reported, a slope of 1 can occur even when calibration is poor.
- This holds true regardless of changes in the definition of calibration over time.

What this adds to what is known?

- Many authors are not following the established advice that slope should not be used to measure calibration alone, but should be accompanied by intercept or other evidence of overall calibration.

What is the implication and what should change now?

- The slope by itself quantifies spread, and the misleading term “calibration slope” should be deprecated to discourage misreporting and misinterpretation.

2. History of the term “calibration slope”

In engineering and chemistry, calibration refers to the process of mapping a reading on an instrument to the actual values to be measured [10]. For example, when reporting an electrochemical method for monitoring blood glucose, Picher et al. reported a “calibration slope dI/dc of 8.7 nA mM⁻¹”: that is, an 8.7 nA higher current corresponds to a 1 mM higher concentration of glucose [11]. Here, calibration refers to a process, and the quantity known as the calibration slope has units.

In clinical epidemiology, calibration refers to a property of a risk score or other numerical prediction rule, and the quantity known as the calibration slope is without units. Recalibration is the process analogous to that in engineering and chemistry, of adjusting an existing model to a new population [8,12]. Methodological publications that are often cited for the calibration slope in clinical epidemiology appear to trace the term to a 1958 paper by Cox that proposed the slope of a (logistic) regression be used to evaluate agreement between a model and a dichotomous outcome [2,13]. In fact, that paper did not use the term “calibration” and referred to the slope as a measure of “spread.” The origins of the term “calibration slope” in epidemiology are therefore unclear, but it may originate with the role of the intercept and slope in the process of recalibration [12].

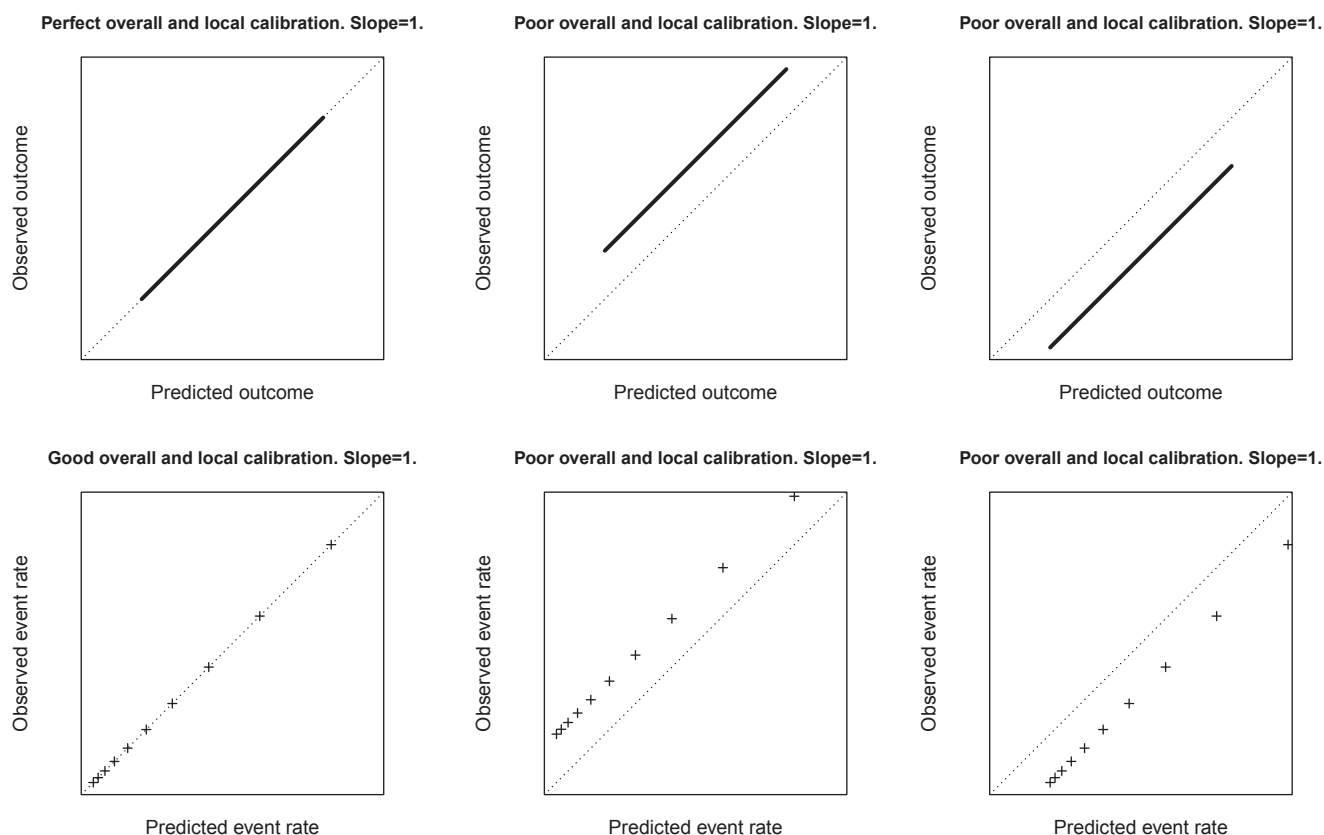


Fig. 1. Six hypothetical models with slope equal to 1. Top row: models for a continuous outcome. Second row: models for a dichotomous outcome.

3. What does the slope measure?

It has been pointed out that the slope does not measure calibration in the sense of overall calibration [14]. In fact, the slope does not by itself measure calibration in any sense: it is recommended for use only in conjunction with the corresponding intercept or other measure of overall calibration (“calibration-in-the-large”) [2,9]. When the intercept is close to 0, a slope close to 1 indicates that good calibration is also maintained across the range of individuals or subgroups, whereas a slope greater than or less than 1 indicates that there are individuals or subgroups in whom calibration is poor.

We note that Cox described it as a measure of “spread” and whether the predicted probabilities “do not vary enough” (slope greater than 1) or “vary too much” (slope less than 1). Some authors have identified spread as an aspect of discrimination [9], but a different aspect of discrimination from that measured by rank/concordance statistics such as the *c*-statistic [15] (page 78).

4. The slope does not by itself measure calibration: illustrative examples

Fig. 1 shows the behavior of the slope for six hypothetical models. In each figure in the top row of Fig. 1, the *x*-axis represents the predicted value of a continuous outcome and the *y*-axis the observed value. The dotted line represents the ideal situation, that predicted outcome equals observed outcome, and the solid line represents the actual situation in some hypothetical case. In the first plot, the solid line overlays the dotted line because the model makes perfect predictions. In the second plot, the solid line is displaced because actual outcomes are higher than those predicted by the model. The model is poorly calibrated, but note that the slope of the line remains 1. Similarly, in the third panel, the calibration is poor but the slope is 1. It will be seen that a slope of 1 can be associated with models that are well or poorly calibrated. This observation holds whether we take calibration to refer to overall calibration, or calibration at a finer level. The examples shown here feature perfect discrimination, in that the ordering of the predictions is identical to the ordering of the outcomes, but it is equally possible to generate equivalent examples with imperfect discrimination.

In each figure of the second row of Fig. 1, the *x*-axis represents the predicted risk, and the *y*-axis the observed rate, of a binary outcome in groups of individuals. Again, this row of figures demonstrates that a slope of 1 may be associated with models that are well or poorly calibrated. Such plots often group individuals by deciles of predicted risk, although more rigorous tests of the model can be based on other groupings [8]. Cox proposed logistic regression of individual binary outcomes on the predicted probabilities [13]. Then it is again possible, though not

shown here, to construct equivalent examples such that the slope remains 1 even as calibration varies.

5. Examples in the clinical literature

The PROOF-BP equation for diastolic blood pressure predicts the difference between blood pressure measured in the clinic and blood pressure measured by a patient at home [16]. In one of the data sets used for external

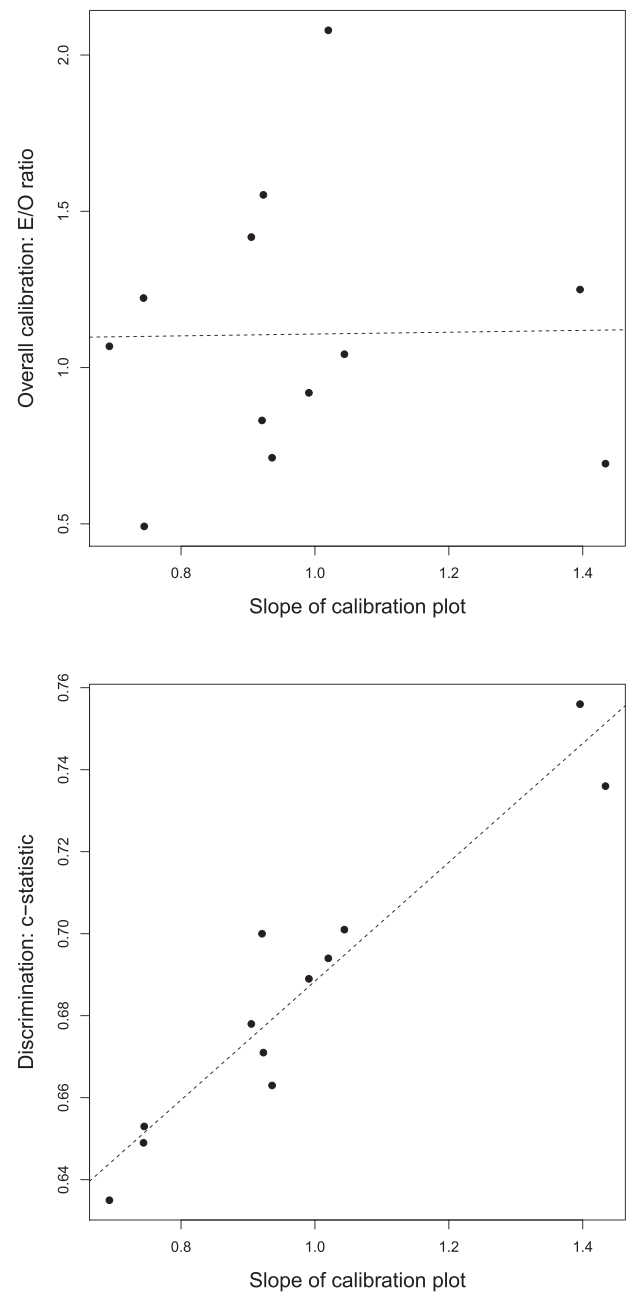


Fig. 2. Relationship of the slope of a calibration plot to (top) calibration in the large, measured by E/O ratio, and (bottom) discrimination, measured by the *c*-statistic, in 12 iterations of cross-validation of a model for deep-vein thrombosis (extracted from Appendix B of [18]). Dotted lines show line of best fit by least squares.

Table 1. Stated interpretation of the “calibration slope” in papers from 2016 to 2017

Author	Type of validation (e.g., internal, external)	Stated interpretation of the calibration slope	Text
External validation			
Basu, Lancet Diabetes & Endocrin, 2017	External	Calibration	“We assessed model calibration through the slope and intercept...” [1]
Besseling, European Heart Journal, 2017	External	Calibration	“... calibration, visualised in a calibration plot, and quantified by the slope of the calibration line”
Dagan, BMJ, 2017	External	Calibration	“To provide calibration measures that are not based on grouping of individuals into strata, we compiled calibration parametric curves, calibration slopes, and calibration-in-the-large values”
Das, Pediatric Blood Cancer, 2017	External	Calibration	“... calibration by the Hosmer-Lemeshow test and calibration slope”
DeMartino, J Vasc Surgery, 2017	External	Calibration	“... the calibration slope is more informative and the preferred method for assessing calibration. A value closer to 1 indicates a better calibration”
Esbenshade, Cancer, 2017	External	Not explicit	“The updated calibration slopes and intercepts were reported”
Haeusler, Br J Cancer, 2017	External	Not explicit	“... performed poorly as reflected by the low AUC and calibration slope”
Mayer, Diab Care, 2017	External	Calibration	“... largely irrelevant deviations from optimal calibration (calibration-in-the-large: 21.125 and 0.95; calibration slopes: 1.07 and 1.13 in the two groups...)”
Nasr, Anes and Analgesia, 2017	External (temporal)	Calibration	“The magnitude of miscalibration was calculated as the calibration slope; the closer the slope coefficient is to 1, the better the calibration”
Ng, Clin Nutr, 2017	External	Calibration	“Calibration of the model was assessed by comparing the predicted mortality with the observed mortality in the development and validation cohorts, using the Hosmer-Lemeshow test and the calibration slope...”
Peters-Sengers, Transplantation, 2017	External	Calibration	“We assessed calibration with the calibration-slope ... A calibration slope > 1 indicates that predicted probabilities do not vary enough, and if < 1 then predicted risks are on average too low for low outcome risks and too high for high outcome risks”
Rogers, Ann Thorac Surg, 2017	External (temporal)	Not explicit	“...an indication of whether the model is underfitting or overfitting the data, and the intercept indicates whether the model is underpredicting or overpredicting deaths. If a model is perfectly calibrated, slope = 1 and intercept = 0”
Thangaratinam, BMC Med, 2017	External	Calibration	“...showed reasonable calibration (slope 0.80).”
Bartels, Neuro-Oncology, 2016	External (temporal)	Discrimination	“The calibration slope was 0.64 + 0.15 (95% CI: 0.34–0.94), indicating poorer discrimination in the validation set than in the derivation set.”
Dean, Int J Radiation Oncology, 2016	External	Calibration	“...[one model] overfitting the data (calibration slope < 1) and [another model] underfitting the data (calibration slope > 1)”
Dhillon, Human Reproduction, 2016	External (temporal)	Calibration	“... perfect calibration is displayed as a straight line passing through zero with a gradient of one.”

(Continued)

Table 1. Continued

Author	Type of validation (e.g., internal, external)	Stated interpretation of the calibration slope	Text
Gerdin, BMC Emergency Medicine, 2016	External (temporal)	Calibration	"Assessed calibration visually using a calibration plot and statistically by estimating the calibration slope."
Inohara, JACC, 2016	External	Calibration	"...slopes and intercepts across deciles of risk... If the model was perfectly calibrated, the intercept and slope would equal 0 and 1, respectively."
Kronisch, Arthritis & Rheumatology, 2016	External	Calibration	"Calibration was examined using ... the calibration slope"
Poyet, Urological Oncology, 2016	External	Calibration	"Calibration and discrimination were assessed using the calibration slope method and the area under the receiver operating characteristic curve (AUC), respectively"
Sarais, Human Reproduction, 2016	External	Not explicit	"The slope of the linear predictor (calibration slope) was..."
Siregar, Circ CV Quality and Outcomes, 2016	External (temporal)	Calibration	"The slope is 1 in a perfectly calibrated model. A calibration slope smaller than 1 indicates that predicted risks were too extreme in the sense of overestimating for patients at high risk while underestimating for patients at low risk and is indicative of overfitting of the model."
Velseboer, Neurology, 2016	External	Calibration	"a well-calibrated model with a calibration slope of 1.13"
Walters, BMC Medicine, 2016	External	Calibration	"The calibration slope suggested good calibration"
Internal validation			
Ahmad et al., PLoS One, 2017	Internal	Not explicit	"predictions made by this model would neither be overestimated nor underestimated"
Nakhjavan-Shahraki, Int J Pediatr, 2017	Internal	Calibration	"...had a proper calibration in prediction of TBI in children (slope = 0.97 and intercept = 0.006)"
Sini, Radiotherapy and Oncology, 2017	Internal	Not explicit	"the performance of the models was evaluated through the calibration plot (slope and regression coefficient R^2)"
Van Walraven, Am J Med, 2017	Internal	Calibration	"measuring overall fit (Regenkirke's R^2), model discrimination (C-statistic), and model calibration (calibration slope)"
Zhang, Oncotarget, 2017	Internal	Not explicit	"Figure 4: Calibration of LASSO score in the validation cohort. The calibration slope was 0.889 and the Brier value was 0.173"
Ballard, Neuropsychologia, 2016	Internal	Calibration.	"calibration of the model was assessed by considering the slope of the predictive index..."
Phillips, Br J Cancer, 2016	Internal-external [19]	Calibration.	"... good calibration (calibration slope 0.95)."
Suemoto, J Gerontol A Biol Sci Med Sci, 2016	Internal	Calibration	"estimated slope between observed and predicted 10-yr mortality risk across deciles of risk (calibration)"
Van Walraven, J Clin Epi, 2016	Internal	Calibration	"The model was ... well calibrated (calibration slope, 1)."

Validation studies using data-splitting methods (including cross-validation and bootstrapping) were classified as "internal" validation if data splitting was carried out at random, or as "external" validation if data splitting was carried out systematically. One study (Phillips, 2016) used a leave-one-out internal-external validation procedure [19].

validation [17], the slope of the calibration plot was 1.015. For illustrative purposes, we created an alternative equation by adding 30 mm Hg to the intercept in the PROOF-BP equation, changing the intercept from -6.98 mm Hg to

23 mm Hg. In the same validation data set, the alternative equation overestimates the difference between home and clinic diastolic blood pressure in all patients (by between 4 mm Hg and 60 mm Hg) and therefore has poor calibration

by any definition. However, the slope of the calibration plot of this new rule remains 1.015.

Given data sets from 12 studies of deep vein thrombosis, Snell et al. used 11 data sets to develop a prediction rule and the 12th to validate it; they then repeated this procedure 12 times, cycling the data sets, so that each data set in turn was the external validation data set [18]. Across the 12 analyses, the slope of the calibration plot showed little correlation with the overall calibration measured by E/O ratio (Fig. 2, top; $r = 0.016$) or by intercept of the calibration plot (figure not shown; $r = -0.059$) but was strongly correlated with the aspect of discrimination measured by the c -statistic (Fig. 2, bottom; $r = 0.95$).

6. Recent use in the literature

Using the Scopus database and limiting to the subject area of Medicine, we identified 40 papers from 2016 to 2017 that used the term “calibration slope” (search using terms TITLE-ABS-KEY [“calibration slope”] AND [LIMIT-TO {SUBJAREA, “MEDI”}] AND [LIMIT-TO {PUBYEAR, 2016} OR LIMIT-TO {PUBYEAR, 2017}] conducted on August 1, 2018). After removing seven methodological papers, 33 validation papers remained. We examined the text for evidence of the authors’ interpretation of the “calibration slope” (Table 1). Twenty-five of the 33 papers explicitly interpreted the slope as a measure of calibration. One paper explicitly interpreted the slope as a measure of discrimination [20]. The remaining seven papers used the term “calibration slope” without explicitly interpreting it as “calibration” or “discrimination.”

Of the 33 papers using “calibration slope” in the Table, 9 also used the intercept, 6 used an alternative measure of overall calibration, and 17 used the slope as the sole measure of calibration. One paper estimated the slope by constraining the intercept to be zero.

7. Why does this matter?

The methodological texts are clear and the slope should be used together with the intercept or (especially for Cox models, which have no intercept) other measure of overall calibration [2,9,15], but as seen in the Table, this advice is often not followed. In almost half of these papers, the slope was the sole attempt to report calibration, and as our examples emphasize, the slope by itself does not measure calibration. We believe that the term “calibration slope,” for a statistic that measures spread but can be unchanged by important changes in calibration, is unhelpful. We argue that as long as this slope is referred to as the “calibration slope,” authors, journals, and readers will be at risk of misreporting and hence of mistaking poorly calibrated models for good models.

When the only validation being conducted is internal validation (evaluation of the model in the same data on which it was developed), then overall calibration is guaranteed to be good for some statistical procedures (including linear and logistic regression). This may not hold for other model-fitting procedures (e.g., neural networks [21]) and in general good calibration on the training set should be regarded as a property to be proven, or evaluated numerically. For validation procedures based on data splitting, especially nonrandom data splitting (e.g., temporal or geographic validation [19]), and for external validation, fully quantifying calibration is always an essential part of assessment of model generalizability.

8. Conclusion and recommendations

We believe that ideally, the term “calibration slope” should be avoided in clinical epidemiology and reserved for its original meaning: a non-unitless slope in applications such as engineering and chemistry. We would prefer to see the unitless slope proposed by Cox in a logistic regression context, and its unitless generalizations in clinical epidemiology, referred to by another name—such as “the Cox slope” or “Cox’s measure of spread.” In practice, however, the term “calibration slope” is deeply embedded in the clinical epidemiology literature. A conservative step would be to encourage authors to use where possible more explicit alternatives such as “the unitless slope of our calibration plot” or “the coefficient of our logistic calibration analysis”—though these are inconveniently verbose. We are encouraged by increasingly sophisticated approaches to evaluation of calibration in the recent literature [22,23]. The most important recommendation is to reiterate that the slope (or intercept) never be reported as a sole measure of calibration [9]. Even in internal validation, we believe that good overall calibration is something to be demonstrated rather than assumed. Desirable properties that are guaranteed, for procedures such as regression that have known statistical properties, cannot be taken for granted in the increasingly wide variety of algorithms labeled as “machine learning” or “artificial intelligence.”

CRediT authorship contribution statement

Richard J. Stevens: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Writing - original draft. **Katrina K. Poppe:** Investigation, Data curation, Writing - review & editing.

Acknowledgments

The authors are grateful to an anonymous reviewer whose comments and suggestions have substantially improved the paper.

Richard Stevens was part-funded by the NIHR Oxford Biomedical Research Centre. Katrina Poppe is supported by the Heart Foundation Hynds Senior Fellowship.

References

- [1] Wallace E, Smith SM, Perera-Salazar R, Vaucher P, McCowan C, Collins G, et al. Framework for the impact analysis and implementation of clinical prediction rules (CPRs). *BMC Med Inform Decis Mak* 2011;11:62.
- [2] Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York, London: Springer; 2009.
- [3] Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–87.
- [4] Miller ME, Hui SL, Tierney WM. Validation techniques for logistic regression models. *Stat Med* 1991;10:1213–26.
- [5] Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453–73.
- [6] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol* 2013;13:33.
- [7] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010;21:128–38.
- [8] 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;74:167–76.
- [9] Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–31.
- [10] Schaschke C. A dictionary of chemical engineering. Oxford, UK: Oxford University Press; 2014.
- [11] Picher MM, Küpcü S, Huang C-J, Dostalek J. Nanobiotechnology advanced antifouling surfaces for the continuous electrochemical monitoring of glucose in whole blood using a lab-on-a-chip. *Lab Chip* 2013;13:1780–9.
- [12] Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJC, Habbema JDF. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004;23:2567–86.
- [13] Cox DR. Two further applications of a model for binary regression. *Biometrika* 1958;45:562–5.
- [14] Vach W. Calibration of clinical prediction rules does not just assess bias. *J Clin Epidemiol* 2013;66:1296–301.
- [15] Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York, London: Springer; 2001.
- [16] Sheppard JP, Stevens R, Gill P, Martin U, Godwin M, Hanley J, et al. Predicting out-of-office blood pressure in the clinic (PROOF-BP): derivation and validation of a tool to improve the accuracy of blood pressure measurement in clinical practice. *Hypertension* 2016;67:941–50.
- [17] McKinsty B, Hanley J, Wild S, Pagliari C, Paterson M, Lewis S, et al. Telemonitoring based service redesign for the management of uncontrolled hypertension: multicentre randomised controlled trial. *BMJ* 2013;346:f3030.
- [18] Snell KIE, Hua H, Debray TPA, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. *J Clin Epidemiol* 2016;69:40–50. (Supplementary materials 3a).
- [19] Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J Clin Epidemiol* 2016;79:76–85.
- [20] Bartels RHMA, De Ruiter G, Feuth T, Arts MP. Prediction of life expectancy in patients with spinal epidural metastasis. *Neuro Oncol* 2016;18:114–8.
- [21] Karhade AV, Thio Q, Ogink P, Kim J, Lozano-Calderon S, Raskin K, et al. Development of machine learning algorithms for prediction of 5-year spinal chordoma survival. *World Neurosurg* 2018;119:e842–7.
- [22] Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Stat Med* 2014;33:517–35.
- [23] Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. *Stat Med* 2019;38:4051–65.