BMC Medicine

OPINION

Open Access

There is no such thing as a validated prediction model

Ben Van Calster[1,2,3], Ewout W. Steyerberg[1], Laure Wynants[1,2,4] and Maarten van Smeden[5*]

Abstract

Background Clinical prediction models should be validated before implementation in clinical practice. But is favorable performance at internal validation or one external validation sufficient to claim that a prediction model works well in the intended clinical context?

Main body We argue to the contrary because (1) patient populations vary, (2) measurement procedures vary, and (3) populations and measurements change over time. Hence, we have to expect heterogeneity in model performance between locations and settings, and across time. It follows that prediction models are never truly validated. This does not imply that validation is not important. Rather, the current focus on developing new models should shift to a focus on more extensive, well-conducted, and well-reported validation studies of promising models.

Conclusion Principled validation strategies are needed to understand and quantify heterogeneity, monitor performance over time, and update prediction models when appropriate. Such strategies will help to ensure that prediction models stay up-to-date and safe to support clinical decision-making.

Keywords Risk prediction models, Predictive analytics, Internal validation, External validation, Heterogeneity, Model performance, Calibration, Discrimination

Background

Clinical prediction models combine multiple patient and disease characteristics to estimate diagnostic or prognostic outcomes. Such models emerge continuously across a broad range of medical fields, often with the goal to guide patient risk stratification and to assist in making optimal decisions for individual patients. Prediction models need validation before implementation in clinical practice [1–3]. *Internal validation* refers to the validation of the model on the same patient population on which it has been developed, for example using a train-test split, cross-validation, or bootstrapping [4]. Conversely, *external validation* refers to the validation of the model on a new set of patients, usually collected at the same location at a different point in time (temporal validation) or collected at a different location (geographic validation) [5, 6].

Whereas internal validation focuses on reproducibility and overfitting, external validation focuses on transportability. Although assessing transportability of model performance is vital, an external validation with favorable performance does not prove universal applicability and does not justify the claim that the model is 'externally valid'. Instead, the aim should be to assess performance across many locations and over time, in order to maximize the understanding of model transportability.

*Correspondence:
Maarten van Smeden
M.vanSmeden@umcutrecht.nl
[1] Department of Development and Regeneration, KU Leuven, Leuven, Belgium
[2] EPI-Center, KU Leuven, Leuven, Belgium
[3] Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands
[4] Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, Netherlands
[5] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, Netherlands

Nevertheless, we argue that it is impossible to definitively claim that a model is 'externally valid', and that such terminology should be avoided. We discuss three reasons for this argument.

Reason 1: patient populations vary

Description

When validating a prediction model on an external dataset, patient characteristics are likely to be different than the characteristics of patients used for model development, even if patients in the validation dataset satisfy the same inclusion and exclusion criteria. Healthcare systems include different types of hospitals or practices, and healthcare systems vary between countries or even regions. Therefore, notable differences in patient characteristics (such as demographics, risk factors, and disease severity) between centers with similar inclusion and exclusion criteria are the rule rather than the exception [7, 8]. Such differences tend to be larger between different types of centers (e.g., secondary versus tertiary care hospitals), or if the validation data uses different inclusion and exclusion criteria. A prediction model that was developed in a tertiary care hospital may yield risk estimates that are invalid for the typical population seen at a regional hospital, or even for tertiary care hospitals in another country [9].

Patient characteristics may not only vary on average, but also in their distribution. Populations with more homogeneous distributions (i.e., less dispersed) tend to have lower discrimination performance as measured for example by the c-statistic or area under the receiver operating characteristic curve. This is because in populations where patients are more alike, estimated risks from the model will also be more alike: it becomes harder to separate those at higher risk from those at lower risk [10].

Besides the discriminative performance, the calibration performance is a key factor in the validity of prediction models. Calibration refers to the agreement between estimated risks from the prediction model and the corresponding observed proportions of events. A common miscalibration situation is that these estimated risks are too high or too low on average (poor "calibration in the large"). Furthermore, estimated risks may be too extreme (too close to 0 or 1) or not extreme enough (too far away from 0 or 1) compared to observed proportions [9]. Miscalibration can be detrimental to the medical decisions that are based on clinical prediction models [11, 12]. For example, if you would like to suggest biopsy when the risk of high-grade prostate cancer is at least 10%, you will perform many unnecessary biopsies when using a risk model that overestimates the risk. Hence, poor calibration is the Achilles heel for applicability of prediction models [9].

Examples

A recent multicenter cohort study externally validated prediction models to diagnose ovarian cancer [13]. Patients with an ovarian tumor were recruited at 17 centers in 7 countries. Participating sites were classified as oncology centers (gynecologic oncology unit within a tertiary center) versus other centers. The mean patient age in the 9 largest centers ($n \geq 166$) varied between 43 and 56 years, and the standard deviation varied between 14 and 19 years (Fig. 1). The median maximum lesion diameter varied between 49 and 70 mm; the interquartile range varied between 38 and 62 mm (Fig. 2). If we focus at oncology centers in Italy (top row in Figs. 1 and 2) in order to compare similar centers from the same country, we still observe different distributions for these variables. Across the whole study sample, 26% of patients at oncology centers had a malignant tumor versus 10% at other centers. All models had higher c-statistics in oncology centers (c-statistics varied between 0.90 and 0.95) versus other centers (0.85 and 0.93).

The Wang clinical model for in-hospital mortality in coronavirus disease 2019 patients was validated using individual participant data from 24 cohorts covering 16 countries [14]. Median cohort size was 283 (range 25 to 25,056), mean patient age varied between 45 and 71 years, the percentage of male patients varied between 45 and 74%. Pooled performance estimates were 0.77 for the c-statistic, 0.65 for the observed over expected (O:E) ratio, and 0.50 for the calibration slope. The O:E ratio < 1 suggests that the model tends to overestimate the risk of in-hospital mortality. The calibration slope < 1 suggests that risk estimates also tend to be too extreme (i.e., too close to 0 or 1). Large heterogeneity in performance was observed, with 95% prediction intervals of 0.63 to 0.87 for the c-statistic, of 0.23 to 1.89 for the O:E ratio, and of 0.34 to 0.66 for the calibration slope. 95% prediction intervals indicate the performance that can be expected when evaluating the model in new clusters.

An external validation study of 104 prediction models for cardiovascular disease reported a median c-statistic of 0.76 for the models in their development data, compared to 0.64 at external validation [12]. When adjusting for differences in patient characteristics, the median c-statistic increased to 0.68. This suggests that about one third of the decrease in discrimination at external validation was due to more homogenous patient samples. This might be expected, given that clinical trial datasets were used for external validation, which often contain more homogeneous samples than observational datasets.

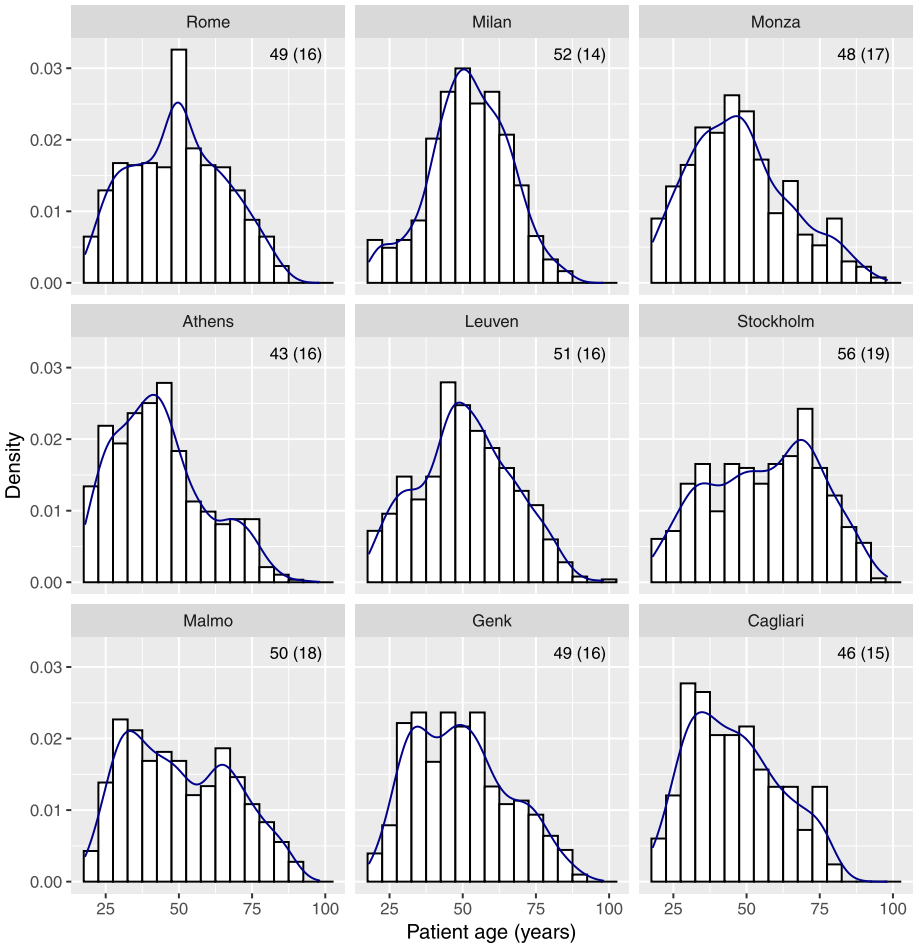

Fig. 1 Distribution of patient age in the 9 largest centers from the ovarian cancer study. Histograms, density estimates, and mean (standard deviation) are given per center

Reason 2: measurements of predictors or outcomes vary

Description

Predictor and outcome measurements or definitions may vary for various reasons, distorting their meaning in a model. First, measurements may be done using equipment from different manufacturers, with different specifications and characteristics. Typical examples are assay kits to quantify biomarker expression, or scanners used to obtain medical images. Second, measurements may depend on a specific method or timing, such as the measurement of blood pressure. Third, measurements may contain high degrees of subjectivity, such that the experience and background of the clinician plays a prominent role. This may cause variable model performance depending on the individual doing the observation. Fourth, biomarker measurements may contain intra-assay variation, analytical variation, and within-subject biological variation (including cyclical rhythms) [15, 16]. Fifth, clinical practice patterns, such as the timing and type of medication or laboratory test orders, tend to vary between clinicians and geographical locations [17, 18]. Such measurements are increasingly used in prediction modeling studies based on electronic health records.

Such heterogeneity in measurement procedures will affect model performance [19, 20]. Depending on how these measurements differ between development and validation, the discriminative performance and in particular the calibration performance can be severely affected. In contrast to intuition, "better" measurements at validation, e.g., predictors measured under stricter protocols than in the development data, may not lead to improved, but instead to deteriorated performance of the prediction model [19, 20].

Examples

Using 17,587 hip radiographs collected from 6768 patients at multiple sites, a deep learning model was trained to predict hip fracture [21]. The c-statistic on the test set (5970 radiographs from 2256 patients; random

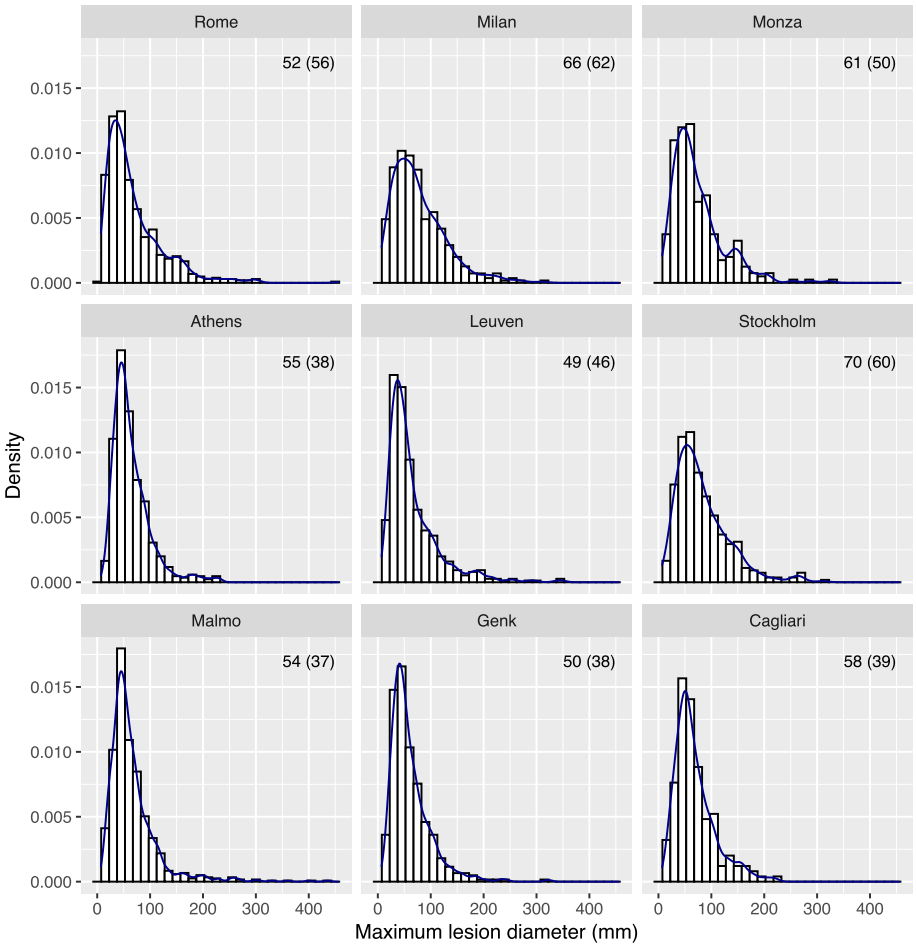

Fig. 2 Distribution of maximum lesion diameter in the 9 largest centers from the ovarian cancer study. Histograms, density estimates, and median (interquartile range) are given per center

train-test split) was 0.78. When non-fracture and fracture test set cases were matched on patient variables (age, gender, body mass index, recent fall, and pain), the c-statistic for hip fracture decreased to 0.67. When matching also included hospital process variables (including scanner model, scanner manufacturer, and order priority), the c-statistic for hip fracture was 0.52. This suggests that variables such as the type of scanner can inflate predictions for hip fracture.

The Wells score calculates the pretest probability of pulmonary embolism in patients suspected to have the condition [22]. A variable in the model is "an alternative diagnosis is less likely than pulmonary embolism". This variable is subjective, and is likely to have interobserver variability. Studies have indeed reported low kappa values for the Wells score (0.38, 0.47) and for the abovementioned subjective variable on its own (0.50) [23, 24].

A systematic review of prognostic models for delirium reported considerable variation in delirium assessment method and frequency across the 27 included studies [25]. Reported methods included the Confusion Assessment Method (CAM), short CAM, Family CAM, Delirium Rating Scale Revised 98, Nursing Delirium Screening Scale, Delirium Assessment Scale, Memorial Delirium Assessment Scale, Delirium Symptom Interview, ward nurse observation, and retrospective chart review. Frequency varied between once to more than once per day. As a result, delirium incidence varied widely.

Seven expert radiologists were asked to label 100 chest x-ray images for the presence of pneumonia [26]. These images were randomly selected after stratification by classification given by a deep learning model (50 images labeled as positive for pneumonia, 50 labeled as negative). There was a complete agreement for 52 cases, 1 deviating label for 24 cases, 2 deviating labels for 13 cases, and 3 deviating labels for 11 experts. Pairwise kappa statistics varied between 0.38 and 0.80, with a median of 0.59.

Wynants and colleagues evaluated the demographic and ultrasound measurements obtained from 2407 patients with an ovarian tumor that underwent surgery

[27]. Each patient was examined by one of 40 different clinicians across 19 hospitals. The researchers calculated the proportion of the variance in the measurements that is attributable to systematic differences between clinicians, after correcting for tumor histology. For the binary variable indicating whether the patient was using hormonal therapy, the analysis suggested that 20% of the variability was attributed to the clinician doing the assessment. The percentage of patients reporting the use of hormonal therapy roughly varied between 0 and 20%. A subsequent survey among clinicians revealed that clinicians reporting high rates of hormonal therapy had assessed this more thoroughly, and that there was a disagreement of the definition of hormonal therapy.

In a retrospective study, 8 radiologists scored four binary magnetic resonance imaging (MRI) features that are predictive of microvascular invasion (MVI) on MRI scans of 100 patients with hepatocellular carcinoma [28]. In addition, the radiologists evaluated the risk of MVI on a five-point scale (definitely positive, probably positive, indeterminate, probably negative, definitely negative). Kappa values were between 0.42 and 0.47 for the features, and 0.24 for the risk of MVI. The c-statistic of the risk for MVI (with histopathology as the reference standard), varied between 0.60 and 0.74.

Reason 3: populations and measurements change over time

Description

Every prediction model is subject to an — usually implicit — expiration date [29]. In the fast-changing and developing world of medicine, patient populations, standards of care, available treatment options, and patient preferences, measurement and data registration procedures change over time [30]. Also, baseline risks for conditions are expected to change over time, for instance, because patient populations tend to become older due to longer life expectancies, or due to shifts in life style and dietary patterns, and the availability of more effective and tailored preventive measures and information. These changes in population characteristics over time are to be expected and may cause performance drifts of prediction models. For example, calibration drift has been well documented [31]. It is therefore increasingly recognized that prediction models need to be updated regularly [32].

A particularly difficult topic is that implementing a prognostic prediction model in clinical practice may invalidate model predictions [33]. The implementations of risk models often aim to identify patients in which interventions are most beneficial. If the implementation of the model leads to effective interventions in high-risk patients, events will be prevented in a proportion of patients. The predictions of the model were derived under the absence of model-induced interventions, and may no longer be accurate; we never observe what could have happened without intervention. In addition, implementation of the model may improve the quality of the measurements of variables that are included as predictors in the model [34]. This should be beneficial as such, but the validity of predictions may be distorted.

Examples

Davis and colleagues developed prediction models for hospital-acquired acute kidney injury using data from patients who were admitted to Department of Veterans Affairs hospitals in the United States in 2003 [35]. The models were developed using different algorithms (e.g., logistic regression, random forest, neural networks), and were validated over time using similar data from patients admitted up to and including 2012. Although discrimination remained fairly stable, with c-statistics roughly around 0.75, there was clear evidence of calibration drift for all models: the risk of the event became increasingly overestimated over time. Accompanying shifts in the patient population were noted: for example, the incidence of the event steadily decreased from 7.7 to 6.2%, age at admission increased, the proportion of patients with a history of cancer or diabetes increased, and the use of various medications increased.

EuroSCORE is a model that predicts in-hospital mortality for patients undergoing cardiac surgery [36]. Using data on 317,292 cardiac surgeries performed in Great Britain and Ireland between 2001 and 2011, it was observed that EuroSCORE overestimated the risk of in-hospital mortality, and that the overestimation aggravated over time [36]. In the beginning of the study period, observed mortality was 4.1% whereas EuroSCORE had an average estimated risk of 5.6%. At the end, observed mortality was 2.8% but the average estimated risk was 7.6%. The c-statistic showed no systematic deterioration, with values varying between 0.79 and 0.85. Furthermore, temporal changes were observed for several predictors (e.g., average age and prevalence of recent myocardial infarction increased) and surgical procedures (e.g., fewer isolated coronary artery bypass graft procedures). The authors further stated that surgeons may have been more willing to operate on patients due to improvements in anesthetic, surgical, and postoperative care.

Conclusions

We presented three reasons why prediction models are never truly validated. A single external validation study in a specific geographical location, in a single time frame, for a sample from a specific patient population is only a snapshot. Such a single study may provide relevant information about the performance of the prediction model

in a specific setting with a particular measurement and time context, but cannot claim transportability beyond that setting. Based on such a study, it is inappropriate to conclude whether a model has been successfully 'validated'. In addition, claims about validity are often based on simplistic criteria using the c-statistic as a measure of discrimination. For example, a model may be declared "validated" if the 95% confidence interval of the c-statistic at validation includes the point estimate of the c-statistic that was originally reported, or if the obtained point estimate of the c-statistic exceeds a certain target value, such as > 0.7 or > 0.8 [37, 38]. Such criteria lack scientific underpinning.

The current focus on developing new models should shift to a focus on more extensive, well-conducted, and well-reported validation studies of promising models. We advise to embrace heterogeneity at model development and at external validation and provide the following general recommendations [10, 39–45].

1) When developing a prediction model, consider the inclusion of multiple settings/locations such as by conducting a multicenter or an individual participant data study [40–42, 45]. Where possible, (a) quantify performance heterogeneity using internal–external cross-validation procedure, where each study is left out once [3, 46]; (b) standardize predictor variables in terms of definition or measurement protocol to reduce prediction measurement heterogeneity; (c) investigate operator-induced measurement variability [27, 47]; and (d) consider to include an operational or population-level characteristic as a predictor (e.g., type of center [45]).
2) When validating a prediction model, inclusion of multiple settings or locations allows to study performance heterogeneity across settings [10, 39, 43, 44].
3) Use appropriate statistical methodology and sample size for model development and/or validation studies, and report fully and transparently [48]. Follow the TRIPOD reporting guideline (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis), including the newly available TRIPOD-Cluster extension where appropriate [49–52]. For example, all predictors should be defined, and the model itself should be made available to allow independent external validation studies [3].
4) Before implementing a model in a specific location, it is recommended to conduct a local validation study. Consider to monitor performance over time and to (dynamically) update the model, in particular when calibration is problematic [31, 32, 53].

We question the requirement from some journals that model development studies should include "an external validation". Apart from the arguments presented above, this requirement may induce selective reporting of a favorable result in a single setting. But is it never good enough? Imagine a model that has been externally validated in tens of locations, representing a wide range of settings, using recent data. Discrimination and calibration results were good, with limited heterogeneity between locations. This would obviously be an important and reassuring finding. Even then, there is still no 100% guarantee that the prediction model will also work fine in a new location. Moreover, it remains unclear how populations change in the future.

In practice, calibration is typically more vulnerable to geographic and temporal heterogeneity than discrimination [9, 12–14, 20, 35, 36, 44]. We stress that calibration assessment in the external validation sample is at least as important as discrimination [9]. If a calibration curve with a narrow 95% confidence interval is close to the ideal diagonal line, one may conclude that risk estimates were appropriate at least for the specific context of the external validation study. For any performance criterion, a meaningful evaluation requires a sufficient sample size. Rules of thumb suggest that at least 100 to 200 cases in the smallest outcome category are required for external validation studies [54, 55]. More refined sample size procedures for model validation have been proposed recently [48].

In conclusion, clinical prediction models are never truly validated due to expected heterogeneity in model performance between locations and settings, and over time. This calls for a stronger focus on validation studies, using principled validation strategies to quantify heterogeneity, regularly monitor model performance, and update models [31, 32]. Such strategies help to ensure that prediction models stay up-to-date to support medical decision-making.

Abbreviations

O:E	Observed over expected
CAM	Confusion Assessment Method
MVI	Microvascular invasion
MRI	Magnetic resonance imaging
TRIPOD	Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Acknowledgements
Not applicable.

Authors' contributions
BVC, EWS, LW, and MvS were involved in study conception. BVC prepared a draft of the manuscript. BVC, EWS, LW, and MvS reviewed and edited the manuscript. BVC, EWS, LW, and MvS approved the final version. All authors agree to take accountability for this work.

Funding
BVC was funded by the Research Foundation – Flanders (FWO; grant G097322N), Internal Funds KU Leuven (grant C24M/20/064), and University Hospitals Leuven (grant COPREDICT). The funders had no role in study design, data collection, data analysis, interpretation of results, or writing of the manuscript.

Availability of data and materials
Information from all examples (except the example on ovarian cancer) was based on information available in published manuscripts. The data used for the ovarian cancer example were not generated in the context of this study and were reused to describe differences between populations from different centers. The data cannot be shared for ethical/privacy reasons.

Declarations

Ethics approval and consent to participate
The reuse of the ovarian cancer data for methodological purposes was approved by the Research Ethics Committee UZ / KU Leuven (number S64709). No consent was required, and the need for individual information letters was waived. Only aggregate data are shown.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 13 October 2022 Accepted: 10 February 2023
Published online: 24 February 2023

References

1. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
2. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245–7.
3. Van Calster B, Wynants L, Timmerman, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc. 2019;26:1651–4.
4. Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol. 2001;54:774–81.
5. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med. 1999;130:515–24.
6. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med. 2000;19:453–73.
7. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. Lancet Digit Health. 2020;2:e489–92.
8. Steyerberg EW, Wiegers E, Sewalt C, Buki A, Citerio G, De Keyser V, et al. Case-mix, care pathways, and outcomes in patients with traumatic brain injury in CENTER-TBI: a European prospective, multicentre, longitudinal, cohort study. Lancet Neurol. 2019;18:923–34.
9. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230.
10. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
11. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making. 2015;35:162–9.
12. Gulati G, Upshaw J, Wessler BS, Brazil RJ, Nelson J, van Klaveren D, et al. Generalizability of Cardiovascular Disease Clinical Prediction Models: 158 Independent External Validations of 104 Unique Models. Circ Cardiovasc Qual Outcomes. 2022;15:e008487.
13. Van Calster B, Valentin L, Froyman W, Landolfo C, Ceusters J, Testa AC, et al. Validation of models to diagnose ovarian cancer in patients managed surgically or conservatively: multicentre cohort study. BMJ. 2020;370:m2614.
14. De Jong VMT, Rousset RZ, Antonio-Villa NE, Buenen AG, Van Calster B, Bello-Chavolla OY, et al. Clinical prediction models for mortality in patients with covid-19: external validation and individual participant data meta-analysis. BMJ. 2022;378:e069881.
15. Ferraro S, Borille S, Carnevale A, Frusciante E, Bassani N, Panteghini M. Verification of the harmonization of human epididymis protein 4 assays. Clin Chem Lab Med. 2016;54:1635–43.
16. White E. Measurement error in biomarkers: sources, assessment and impact on studies. IARC Sci Publ. 2011;163:143–61.
17. Berndt ER, Gibbons RS, Kolotilin A, Taub AL. The heterogeneity of concentrated prescribing behavior: Theory and evidence from antipsychotics. J Health Econ. 2015;40:26–39.
18. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;360:k1479.
19. Luijken K, Groenwold RHH, Van Calster B, Steyerberg EW, van Smeden M. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. Stat Med. 2019;38:3444–59.
20. Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH. Impact of predictor measurement heterogeneity across settings on the performance of prediction models: A measurement error perspective. J Clin Epidemiol. 2020;119:7–18.
21. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. npj Digit Med. 2019;2:31.
22. Wells PS, Anderson DR, Rodger M, Ginsberg JS, Kearon C, Gent M, et al. Derivation of a simple clinical model to categorize patients probability of pulmonary embolism: increasing the models utility with the SimpliRED D-dimer. Thromb Haemostat. 2000;83:416–20.
23. Fesmire FM, Brown MD, Espinosa JA, Shih RD, Silvers SM, Wolf SJ, et al. Critical issues in the evaluation and management of adult patients presenting to the emergency department with suspected pulmonary embolism. Ann Emerg Med. 2011;57:628-652.e75.
24. Iles S, Hodges AM, Darley JR, Frampton C, Epton M, Beckert LEL, et al. Clinical experience and pre-test probability scores in the diagnosis of pulmonary embolism. QJM. 2003;96:211–5.
25. Lindroth H, Bratzke L, Purvis R, Brown R, Coburn M, Mrkobrada M, et al. Systematic review of prediction models for delirium in the older adult inpatient. BMJ Open. 2018;8:e019223.
26. Kim D, Chung J, Choi J, Succi MD, Conklin J, Figueiro Longo MG, et al. Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. Nat Commun. 2022;13:1867.
27. Wynants L, Timmerman D, Bourne T, Van Huffel S, Van Calster B. Screening for data clustering in multicenter studies: the residual intraclass correlation. BMC Med Res Methodol. 2013;13:128.
28. Min JH, Lee MW, Park HS, Lee DH, Park HJ, Lim S, et al. Interobserver Variability and Diagnostic Performance of Gadoxetic Acid-enhanced MRI for Predicting Microvascular Invasion in Hepatocellular Carcinoma. Radiology. 2020;297:573–81.
29. Reynard C, Jenkins D, Martin GP, Kontopantelis E, Body R. Is your clinical prediction model past its sell by date? Emerg Med J. 2022. https://doi.org/10.1136/emermed-2021-212224.
30. Nestor B, McDermott MBA, Boag W, Berner G, Naumann T, Hughes MC, et al. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. Proc Mach Learn Res. 2019;106:1–23.
31. Davis SE, Greevy RA Jr, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. J Biomed Inform. 2020;112:103611.
32. Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagn Progn Res. 2021;5:1.
33 Lenert MC, Matheny ME, Walsh SG. Prediction models will be victims of their own success, unless. J Am Med Inform Assoc. 2019;26:1645–50.
34. Groenwold RHH. Informative missingness in electronic health record systems: the curse of knowing. Diagn Progn Res. 2020;4:8.

35. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24:1052–61.
36. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. Eur J Cardiothorac Surg. 2013;43:1146–52.
37. Cook G, Royle KL, Pawlyn C, Hockaday A, Shah V, Kaiser MF, et al. A clinical prediction model for outcome and therapy delivery in transplant-ineligible patients with myeloma (UK Myeloma Research Alliance Risk Profile): a development and validation study. Lancet Haematol. 2019;6:e154–66.
38. Fan J, Upadhye S, Worster A. Understanding receiver operating characteristic (ROC) curves. CJEM. 2006;8:19–20.
39. Steyerberg EW, Nieboer D, Debray TPA, van Houwelingen HC. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: an overview and illustration. Stat Med. 2019;38:4290–309.
40. Debray TPA, Moons KGM, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med. 2013;32:3158–80.
41. Debray TPA, Damen JAAG, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res. 2019;28:2768–86.
42. Wynants L, Vergouwe Y, Van Huffel S, Timmerman D, Van Calster B. Does ignoring clustering in multicenter data influence the performance of prediction models? A simulation study. Stat Methods Med Res. 2018;27:1723–36.
43. 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.
44. Austin PC, van Klaveren D, Vergouwe Y, Nieboer D, Lee DS, Steyerberg EW. Validation of prediction models: examining temporal and geographic stability of baseline risk and estimated covariate effects. Diagn Progn Res. 2017;1:12.
45. Wynants L, Kent DM, Timmerman D, Lundquist CM, Van Calster B. Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting. Diagn Progn Res. 2019;3:6.
46. Takada T, Nijman S, Denaxas S, Snell KIE, Uijl A, Nguyen TL, et al. Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets. J Clin Epidemiol. 2021;137:83–91.
47. Stiell IG, Clement CM, O'Connor A, Davies B, Leclair C, Sheehan P, et al. Multicentre prospective validation of use of the Canadian C-Spine Rule by triage nurses in the emergency department. CMAJ. 2010;182:1173–9.
48. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. Stat Med. 2021;40:4230–51.
49. 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;68:134–43.
50. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1-73.
51. Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ. 2023;380:e071018.
52. Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data (TRIPOD-Cluster): explanation and elaboration. BMJ. 2023;380:e071058.
53. Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. BMC Med Res Methodol. 2022;22:316.
54. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35:214–26.
55. 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 datra. J Clin Epidemiol. 2016;74:167–76.

Publisher's Note