


COMMENTARY

Open Access



Targeted validation: validating clinical prediction models in their intended population and setting

Matthew Sperrin^{1*} , Richard D. Riley², Gary S. Collins³ and Glen P. Martin¹

Abstract

Clinical prediction models must be appropriately validated before they can be used. While validation studies are sometimes carefully designed to match an intended population/setting of the model, it is common for validation studies to take place with arbitrary datasets, chosen for convenience rather than relevance. We call estimating how well a model performs within the intended population/setting “targeted validation”. Use of this term sharpens the focus on the intended use of a model, which may increase the applicability of developed models, avoid misleading conclusions, and reduce research waste. It also exposes that external validation may not be required when the intended population for the model matches the population used to develop the model; here, a robust internal validation may be sufficient, especially if the development dataset was large.

Keywords: Clinical prediction model, Validation, Generalisability

Background

Clinical prediction models (CPMs) calculate risk of current (diagnostic) and future (prognostic) events for individuals [1]. For example, QRISK calculates 10-year risk of cardiovascular outcomes [2], and EuroSCORE calculates risk of in-hospital mortality following major cardiac surgery [3]. The traditional pipeline for CPM production begins with model development, including internal validation; this is followed by external validations of the model’s performance in different data; the model’s impact may then be tested (e.g., whether its use improves health outcomes), and if considered suitable, the model may be implemented. This pipeline applies equally whether models are developed using AI or machine learning techniques, or regression-based models.

Internal validation is an examination of model performance in the same dataset that was used to develop the CPM. It is important that internal validation corrects for *in-sample optimism*, which is the tendency of models to overfit (perform better in) the development data compared with other data from the same population. This is ideally done using cross-validation or bootstrapping, but is also commonly done by splitting the dataset into training and validation subsets. For example, in the development and internal validation of a prognostic model for muscle injury in elite soccer players, an apparent c-index (a measure of the model’s ability to distinguish cases from non-cases, where a value of 1 is perfect, and 0.5 is no better than chance) of 0.64 reduced to 0.59 when using bootstrapping for optimism adjustment [4].

In contrast, external validation is an examination of model performance in different dataset(s), often regarded as a gold-standard of model ‘credibility’. Selection of the dataset(s) is critical, because model performance is highly dependent on the population and setting [5, 6]. Here, *population* refers to the group of individuals under

*Correspondence: matthew.sperrin@manchester.ac.uk

¹ Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
Full list of author information is available at the end of the article



© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

consideration—e.g., people of a certain age, people in a specific country, people who suffer from a particular disease (and any combinations thereof). *Setting* refers to the place in which the CPM would be used, such as in hospital, primary care, general population, etc. Accordingly, there are at least three types of external validation studies. The first is where researchers investigate model performance in *one particular* population and setting carefully chosen to match the intended use of the model. This might be the same as (or similar to) the population/setting used for model development (assessing reproducibility), or might be a different population/setting (assessing transportability, e.g., evaluating if a model developed for adults has predictive value in children). A second type is where researchers investigate model performance across *multiple* populations and settings, where each is relevant to the intended use (assessing generalisability) [7, 8]; for example, in an individual patient data meta-analysis of performance across multiple countries, with a focus on identifying heterogeneity in performance [5]. In these first two types, the validation dataset(s) match the target population(s) and setting(s) where the CPM is intended for deployment, so the validation is meaningful (provided the methodological quality is also high). A third type is where researchers examine model performance in a new, conveniently available dataset, which is neither representative of the population nor the setting of interest. For example, in a comprehensive review of COVID-19 external validation studies, 35 studies were found to be at high risk of bias in the participant/data domain, which reflects the use of inappropriate dataset(s) for external validation [9]. In these cases, the validation dataset bears little relevance to any target population and setting, and thus the findings have the potential to mislead.

The aim of this paper is to describe why it is necessary to validate a CPM in a population and setting that represents each intended target population and setting of the CPM, and in a manner that reflects each intended use. These populations and settings need to be clearly reported in every validation study. We use the term *targeted validation*, which emphasises that how (and in what data) to validate a CPM should depend on the intended use of the model.

Targeted validation

When a CPM is developed, it should be done so with a clearly defined *intended use* and *population*: i.e., when predictions are to be made, in whom, and for what purpose. Validation should be carried out to show how well the CPM performs at that specific task—a targeted validation. A focus on targeted validation has several advantages. First, a targeted validation study provides estimates

of predictive performance for the intended target setting, so are extremely informative for that setting. Second, the CPM may be (perhaps subsequently) used in many clinical settings and populations—each of which may require its own targeted validation. For example, EuroSCORE was developed to predict risk of in-hospital mortality following major cardiac surgery [3, 10], but validation studies have examined if it could be used in other cardiac surgical interventions [11, 12], i.e., a different population *and* setting. For any given setting, one can assess if existing validation studies sufficiently capture the new intended use(s)/population(s), or if further validations are required. Similarly, where populations that do not match the target are used for validation, the differences can be highlighted as a ‘validation gap’ to be acknowledged or addressed (see “Validation gap” section below). Third, it focuses attention on developing and validating models that have clearly defined uses in practice, since the intended use needs to be defined a-priori, thereby avoiding research waste.

To motivate this idea, consider the following example (see Table 1). A CPM called T-MACS was developed for the prediction of acute myocardial infarction in patients presenting to the emergency department with chest pain [13]. Initially, suppose the intended use of the CPM is to aid clinical decision-making within Hospital A in Manchester, UK. The targeted validation should assess how well the model performs in (a representative sample of) patients from Hospital A, and not how the model generalises to other hospitals [15, 16]. Subsequently, suppose Hospital B in London, UK, wishes to implement the CPM; a new targeted validation should be undertaken to estimate model performance in Hospital B. The CPM has not changed but the intended target population has, hence the required validation is different.

Different targeted validation exercises are important because performance in one target population gives little indication of performance in another [6]. Indeed, performance is likely highly heterogeneous across populations and settings [5], due to differences in case mix (i.e., the distributions of the patient characteristics in the population), baseline risk, and predictor-outcome associations. Therefore, any discussion of validity must be in the context of the target population and setting. It is incorrect to refer to a model as ‘valid’ or ‘validated’ in general—we can only say that a model is ‘valid for’ or ‘validated for’ the particular populations or settings in which this has been assessed. Targeted validation addresses this by *first* identifying the population and setting where a model is intended to be used, and *second* identifying suitable datasets for validations that match the intended population and setting. In addition to avoiding acting on potentially misleading validation studies, a focus on targeted

Table 1 Consider T-MACS—a CPM developed for the prediction of acute myocardial infarction in patients presenting to the emergency department with chest pain [13]. Suppose our intended use is initially for hospitals within the Greater Manchester (UK) area, and then we are considering rolling out the CPM across the UK

External validation where...	Example
...a particular population and setting of intended use is of interest	T-MACS was developed using data from a hospital in Manchester, and validated using data from other hospitals in Manchester [13]. Hence the original development and validation results match the intended use. If, however, we wished to use T-MACS in London, UK, the validation above would be of limited value. We would need a new targeted validation to examine performance in London, UK.
...multiple populations and settings of intended use are of interest	Suppose we wished to implement T-MACS across all hospitals in the UK. Then, validation would be required across many hospitals in the UK, potentially using individual patient data meta-analysis of performance across hospitals to evaluate heterogeneity in performance [5]. Such a validation could also be useful to indicate expected performance of T-MACS in UK hospitals not included in the validation set. This is because, if heterogeneity in performance is low across the included hospitals, this gives some confidence that the CPM will perform well in all areas in the UK.
...an arbitrary dataset is used without consideration of the intended population or setting	A further validation was conducted in hospitals in Australia and New Zealand [14]. For our intended use, this validation offers little evidence. However, it would be very valuable if we were considering using the model in Australia and New Zealand.

validation will also reduce research waste, since being explicit about the target population and use will avoid uninformative studies being conducted. To be concrete, a validation study should not take place unless a population and setting has been identified in which the CPM could potentially be used, and the validation study should be designed to estimate performance in *that* population and setting.

This is not a new idea: we are simply making it explicit. Riley et al. [5], state ‘external validation uses new participant level data, external to those used for model development, to examine whether the model’s predictions are reliable (that is, accurate enough) in individuals from potential population(s) for clinical use’ while Wessler et al. [6], remark that we should not accept a CPM in a particular context ‘unless CPM performance is specifically known to be excellent in populations like those.’ It is also emphasised in the PROBAST risk of bias tool for systematic reviews and meta analyses of CPMs, where the ‘applicability’ domain checks whether included studies consider the same setting and population as the review question [17], emphasised in a recent scoping review of guidance for prediction models using AI [18], and included in the protocol for reporting and risk of bias tools for prediction models developed using AI [19]. *Target validity* in the clinical trial literature, which quantifies bias in transporting a trial-estimated causal effect to a target population, has a similar motivation [20].

Moreover, the targeted validation framework suggests that there may be contexts where the data used for validation could be the same as for development. In the first part of the example above, the closest data to

the intended target population (Hospital A) may be the development data (Table 1). In this case, there is little to be gained through evaluating performance in other hospitals; instead, the focus should be on a thorough internal validation using the development dataset. This internal validation is likely to give a robust estimate of the model’s performance when appropriate steps were taken during the model development to ensure the study has adequate sample size [21], that overfitting is minimised [22], that in-sample optimism is estimated precisely and corrected for [8], and that the optimism was examined by replaying all the model development steps. Moreover, the internal validation should include, for example, temporal or demographic subgroups to test the reproducibility and generalisability of the model. Provided all these steps are thoughtfully conducted, internal validation can be viewed as a reliable measure of performance in the intended population, and the lack of any external validation is not a concern. Indeed, whenever a new model is developed, the model development data should always be chosen according to the anticipated target population and setting: for example, if a model is intended to be used in UK primary care, then UK primary care data should be used to develop the model.

One size fits all versus tailored models

Consider the situation where we wish to implement T-MACS across all hospitals in the UK (Table 1). Here, we could evaluate the CPM in each hospital, and then—depending on the observed performance, and a subsequent impact assessment study—choose to either deploy the model as originally specified, or deploy it after

updating it for each particular context [23, 24]. This situation—in which one wishes to implement a CPM across multiple populations/settings—is common, and there are two main ways of achieving this: building a single CPM for use in all target settings or building tailored CPMs for each target setting.

Under the first approach, one needs to assess generalisability of the CPM [25–27]. A natural way of doing this is to obtain (new) datasets from multiple populations (e.g., across countries, or across clusters of data within electronic health records), evaluate performance of the relevant model in each dataset/cluster, and then meta-analyse [28], with particular attention to quantifying and identifying sources of heterogeneity [5]. Alternatively, we might conduct internal-external cross-validation to combine model development with assessment of model generalisability across the multiple populations/settings [5, 8, 29].

However, developing a model that generalises across multiple populations is difficult, not least because predicted risks are unlikely to calibrate well with observed risks in every population and setting. Methods are emerging that support this [30], and incorporating causal inference principles is also likely to help generalisability and transportability [31, 32]. Nevertheless, insisting on a model with broad general applicability comes at the price of reduced performance in specific settings or populations [15]. Model performance being worse in specific subgroups also raises concerns over fairness [33]. As such, the second approach—in which one starts with a CPM developed using sufficient (and appropriate) data [21], and then tailors or updates it to local settings [23, 24]—may be appealing. This implies *targeted updating* of a given CPM updated for specific target populations/settings; following this, targeted validation exercises would be needed in each local population/setting to examine the locally tailored CPM. However, the feasibility of having a large family of tailored CPMs is a challenge with regard to provenance and maintenance.

Validation gap

Focus on targeted validation makes the interpretation of the predictive performance clearer. If the target population is patients in Hospital B, then we need to estimate model performance in Hospital B. If we can obtain data, and have the resource, to validate the model in this population, then the corresponding performance estimates are appropriate. However, if the validation had instead been performed in Hospital C (for example, if there is little or inadequate historical data available in Hospital B, or resource constraints do not permit the validation study to be conducted in Hospital B), then

targeted validation allows one to infer how applicable the predictive performance estimates we obtain might be for Hospital B given the difference between the two settings: a ‘validation gap’. Identification of a validation gap suggests caution in using the CPM within the target population. In this situation, we recommend that differences between the validation population and target population can be described qualitatively, such as by contrasting the setting, case-mix and patient eligibility criteria; or quantitatively (where sufficient data exists), by examining membership models for whether individuals belong to the validation population or target population [25, 34]. We then recommend being explicit about the required assumptions for the validation results to transport, and to address the differences where possible, such as through reweighting the validation population to resemble the target population [35]. This reweighting could be done at individual level, or at group level—for example, if performance is known to vary across groups of patients in different disease subgroups, then performance in each of the subgroups could be reported in the validation population, then combined through appropriate weighting to estimate performance in the target population. Such reweighting would allow estimation of global performance measures such as AUC, while the weights themselves can be used to infer where the differences between the validation and target are largest (e.g., under-represented subgroups), and therefore where the CPM may have poor local performance in the target population (i.e., issues with strong calibration as defined in [36]). The validation gap concept can also help researchers to decide when a full new validation exercise might not be necessary—i.e., where existing validations are performed in sufficiently similar populations and settings, and the model has been shown to be generalisable.

A particular challenge that targeted validation emphasises is that the implementation of the CPM will always be *after* validation and subsequent impact study—so a validation gap in time will always be present [37]. CPMs can be prone to changes in the underlying distribution, which causes calibration drift and other performance issues, particularly in contexts such as surgical risk [38], and infectious disease risk [39]. Therefore, model development strategies that allow a CPM to respond to changes over time—such as dynamic modelling [40, 41] or temporal recalibration [42]—are very promising. This also emphasises the importance of a validation exercise thoroughly considering heterogeneity in geography, over time, and over setting [5], and indeed the possibility of targeted updating, in which CPMs are updated to a new time period before revalidating.

Conclusion

We recommend that validation of clinical prediction models should relate to the target population and setting, and suggest using the term targeted validation to make this focus explicit. This provides a framework in which researchers are transparent about the intended use of the model being validated, and motivates the use of a validation dataset that is representative of the population(s) and setting(s) of intended use. If such a dataset is not available, and validation is undertaken, then researchers should highlight differences between the validation and target populations (a ‘validation gap’) so that the findings can be placed in context.

There are three key implications of focusing on targeted validation. First, validation studies that do not state, and clearly justify, the intended target population or setting are not fit for purpose. The prevalence of this problem has not yet been quantified. Second, a new intended use of a model requires a new targeted validation exercise (which may be a new validation, or careful consideration of the relevance of existing validations): CPMs should not be referred to as ‘valid’ or ‘validated’ as this is meaningless without reference to a target population. Third, external validation studies may not always be needed, specifically if the development dataset is sufficiently large, already represents the target population and setting, and appropriate steps have been taken to adjust performance estimates for in-sample optimism.

Abbreviations

CPM: Clinical prediction model; EuroSCORE: European system for cardiac operative risk evaluation; T-MACS: Troponin-only Manchester acute coronary syndromes decision aid.

Acknowledgements

Not applicable.

Authors' contributions

MS wrote the draft. RDR, GSC, and GPM reviewed and refined the draft, and all authors read and approved the final article.

Funding

GSC was supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (programme grant: C49297/A27294). RDR and GSC were supported by funding from the MRC Better Methods Better Research panel (grant reference: MR/V038168/1). GPM and RDR were supported by funding from the MRC-NIHR Methodology Research Programme (grant number: MR/T025085/1).

Availability of data and materials

No data were generated or analysed in support of this research.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Division of Informatics, Imaging and Data Science, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. ²Institute of Applied Health Research, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. ³Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK.

Received: 24 August 2022 Accepted: 14 November 2022

Published online: 22 December 2022

References

1. Steyerberg EW. Clinical prediction models : a practical approach to development, validation, and updating. New York: Springer; 2019. p. 497.
2. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*. 2017;357:j2099.
3. Nashef SAM, Roques F, Sharples LD, Nilsson J, Smith C, Goldstone AR, et al. EuroSCORE II. *Eur J Cardiothorac Surg*. 2012;41(4):734–45.
4. Hughes T, Riley RD, Callaghan MJ, Sergeant JC. The value of preseason screening for injury prediction: the development and internal validation of a multivariable prognostic model to predict indirect muscle injury risk in elite football (soccer) players. *Sports Med - Open*. 2020;6(1):22.
5. 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.
6. Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, et al. External validations of cardiovascular clinical prediction models: a large-scale review of the literature. *Circ Cardiovasc Qual Outcomes*. 2021;14(8):e007858.
7. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med*. 1999;130(6):515–24.
8. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245–7.
9. Wynants L, Van Calster B, Bonten MM, Collins GS, Debray TP, De Vos M, et al. Systematic review and critical appraisal of prediction models for diagnosis and prognosis of COVID-19 infection. *BMJ*. 2020. <https://doi.org/10.1101/2020.03.24.20041020>.
10. Nashef SAM, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R, et al. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg*. 1999;16(1):9–13.
11. Martin GP, Sperrin M, Ludman PF, de MA B, Gale CP, Toff WD, et al. Inadequacy of existing clinical prediction models for predicting mortality after transcatheter aortic valve implantation. *Am Heart J*. 2017;184:97–105.
12. Durand E, Borz B, Godin M, Tron C, Litzler PY, Bessou JP, et al. Performance analysis of EuroSCORE II compared to the original logistic EuroSCORE and STS scores for predicting 30-day mortality after transcatheter aortic valve replacement. *Am J Cardiol*. 2013;111(6):891–7.
13. Body R, Carlton E, Sperrin M, Lewis PS, Burrows G, Carley S, et al. Troponin-only Manchester Acute Coronary Syndromes (T-MACS) decision aid: single biomarker re-derivation and external validation in three cohorts. *Emerg Med J*. 2017;34(6):349–56.
14. Greenslade JH, Nayer R, Parsonage W, Doig S, Young J, Pickering JW, et al. Validating the Manchester Acute Coronary Syndromes (MACS) and Troponin-only Manchester Acute Coronary Syndromes (T-MACS) rules for the prediction of acute myocardial infarction in patients presenting to the emergency department with chest pain. *Emerg Med J*. 2017;34(8):517–23.
15. 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(9):e489–92.
16. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? *Clin Kidney J*. 2021;14(1):49–58.

17. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51–8.
18. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *Npj Digit Med*. 2022;5(1):1–13.
19. Collins GS, Dhiman P, Navarro CLA, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
20. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target validity and the hierarchy of study designs. *Am J Epidemiol*. 2019;188(2):438–43.
21. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
22. Martin GP, Riley RD, Collins GS, Sperrin M. Developing clinical prediction models when adhering to minimum sample size recommendations: the importance of quantifying bootstrap variability in tuning parameters and predictive performance. *Stat Methods Med Res*. 2021;30(12):2545–61.
23. Janssen K, Moons K, Kalkman C, Grobbee D, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76–86.
24. Steyerberg EW, Borsboom GJJM, 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(16):2567–86.
25. 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):279–89.
26. Toll D, Janssen K, Vergouwe Y, Moons K, et al. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. 2008;61(11):1085–94.
27. Cabitza F, Campagner A, Soares F, García de Guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput Methods Programs Biomed*. 2021;208:106288.
28. Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res*. 2018;27(11):3505–22.
29. 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;0(0) Available from: [https://www.jclinepi.com/article/S0895-4356\(21\)00107-4/abstract](https://www.jclinepi.com/article/S0895-4356(21)00107-4/abstract).
30. de Jong VMT, KGM M, MJC E, Riley RD, TPA D. Developing more generalizable prediction models from pooled studies and large clustered data sets. *Stat Med*. n/a(n/a). Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8981>.
31. Sperrin M, Díaz-Ordaz K, Pajouheshnia R. Invited Commentary: Treatment drop-in: making the case for causal prediction. *Am J Epidemiol*. 2021;190(10):2015–8.
32. Bellamy D, Hernán MA, Beam A. A structural characterization of shortcut features for prediction. *Eur J Epidemiol*. 2022;37(6):563–8.
33. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *Npj Digit Med*. 2020;3(1):1–8.
34. Schat E, van de Schoot R, Kouw WM, Veen D, Mendrik AM. The data representativeness criterion: predicting the performance of supervised classification based on data set similarity. *Zhang J, PLoS One*. 2020;15(8):e0237009.
35. Riley RD, Tierney J, Stewart LA (Eds). *Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research*. Chichester: Wiley; 2021.
36. 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.
37. Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Appl Clin Inform*. 2021;12(4):808–15.
38. 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*. 2012;43(6):1146–52.
39. Clift AK, Coupland CA, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, et al. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ*. 2020;371.
40. Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagn Progn Res*. 2018;2(1):23.
41. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform*. 2020;112:103611.
42. Booth S, Riley RD, Ensor J, Lambert PC, Rutherford MJ. Temporal recalibration for improving prognostic model development and risk predictions in settings where survival is improving over time. *Int J Epidemiol*. 2020; Available from: <https://academic.oup.com/ije/advance-article/doi/10.1093/ije/dyaa030/5815624>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

