Perspectives on validation of clinical predictive algorithms

Hond, A.A.H. de; Shah, V.B.; Kant, I.M.J.; Calster, B. van; Steyerberg, E.W.; Hernandez-Boussard, T.

npj | digital medicine

www.nature.com/npjdigitalmed

COMMENT OPEN

Check for updates

Perspectives on validation of clinical predictive algorithms

Anne A. H. de Hond[1,2,3]✉, Vaibhavi B. Shah[2], Ilse M. J. Kant[4], Ben Van Calster[3,5], Ewout W. Steyerberg[1,3] and Tina Hernandez-Boussard[2,6,7]

The generalizability of predictive algorithms is of key relevance to application in clinical practice. We provide an overview of three types of generalizability, based on existing literature: temporal, geographical, and domain generalizability. These generalizability types are linked to their associated goals, methodology, and stakeholders.

npj Digital Medicine (2023)6:86 ; https://doi.org/10.1038/s41746-023-00832-9

Machine learning has led to a surge in the development of clinical predictive algorithms. The generalizability of these algorithms often goes untested[1], leaving the community in the dark on their accuracy and safety when applied to a specific medical setting. We need clear objectives with respect to generalizability that align with the intended use. Journals, funding organizations, and regulatory bodies provide some guidance on generalizability requirements for clinical predictive algorithms, but a clear definition is often lacking. For example, it is considered best practice to 'Describe the generalizability of the model including the performance of the model on validation and testing datasets[2]'. We consider this recommendation too vague. It is not clear what type of generalizability is referred to and whether it is sufficient for the intended use of the algorithm (see Supplementary Table 1 for more examples and suggestions for improvement). This commentary aims to provide clarity on different objectives related to generalizability via an overview of three main types of generalizability summarized from the literature with their associated goals, methodology, and stakeholders.

We performed a scoping review to identify different types of generalizability (see Supplementary Methods and Supplementary Table 2). In the context of clinical prediction models or predictive algorithms, generalizability refers to an algorithm's ability to perform adequately across different settings[3]. Setting is defined by the clinical context of included patients, time, and place. Algorithm performance can then be assessed along various axes, including discrimination[3], calibration[4], and measures for clinical usefulness, such as Net Benefit[5]. We extracted three distinct types of generalizability. Examples of published validation use cases for each generalizability type can be found in Supplementary Table 3.

A key distinction should be made between internal and external validation (Fig. 1). Internal validation assesses the reproducibility of algorithm performance in data that is distinct from the development (or: train) data but derived from the exact same underlying population. It provides an optimism-corrected estimate of performance for the setting where the data originated from[6]. Cross-validation and bootstrapping are the recommended methods to assess internal validity[6,7]. Cross-validation splits the data in equal parts (usually five or ten) and trains the algorithm on all but one holdout part that is used for testing. This process is repeated until all parts have been used as test data. The whole procedure is preferably repeated multiple times for more stability, e.g., a 10×10-fold cross-validation procedure. Bootstrapping repeatedly samples data points from the development data with replacement (usually 500–2000 times). These samples are used to train the algorithm with the original development data as test set[6,8]. Internal validation is necessary but not sufficient to ensure safe clinical applicability. The main stakeholder is the developer of the algorithm, who uses internal validation to assess the validity of the development process, and quantifies overoptimism in expected performance[7,9].

External validity assesses the transportability of the clinical predictive algorithm to other settings than those considered during development (Fig. 1). It encompasses three generalizability types: temporal, geographical and domain generalizability. Temporal validity assesses the performance of an algorithm over time at the development setting. This type of generalizability is required to understand data drift (a change in the data over time from the data that was used during development)[10]. Temporal validity may be assessed by testing the algorithm on a dataset derived from the same setting as the development cohort but from a later time. Variations in design are possible, such as a 'waterfall' design, in which the development time window is repeatedly increased[11]. The main stakeholders of temporal validity are clinicians, hospital administrators, and other clinical end-users that plan to implement the algorithm into their clinical practice. These stakeholders need proof of temporal validity to ensure the safe use of the algorithm at their local clinical institution or hospital.

Geographical validation assesses the generalizability of an algorithm to a place (institution or location) that is different from where the algorithm was developed. This type of validation assesses the heterogeneity across places. Geographical validity can be assessed by testing the algorithm on data collected from the new place(s). More complex designs are possible, such as a leave-one-site-out (or internal-external) validation in which the algorithm is developed on all but one location and tested on the left-out one[12]. This process is repeated until all locations have been used as test location. Geographical validation is required when the algorithm is going to be used outside of the original development place. The main stakeholders are the clinical end-users at a new implementation site who want proof of validity for safe use at their site. Manufacturers, insurers, and governing bodies could be other stakeholders that are interested in

[1]Clinical AI Implementation and Research Lab, Leiden University Medical Centre, Leiden, the Netherlands. [2]Department of Medicine (Biomedical Informatics), Stanford University, Stanford, CA, USA. [3]Department of Biomedical Data Sciences, Leiden University Medical Centre, Leiden, the Netherlands. [4]Department of Digital Health, University Medical Center Utrecht, Utrecht, the Netherlands. [5]Department of Development & Regeneration, KU Leuven, Leuven, Belgium. [6]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [7]Department of Epidemiology & Population Health (by courtesy), Stanford University, Stanford, CA, USA. ✉email: a.a.h.de_hond@lumc.nl

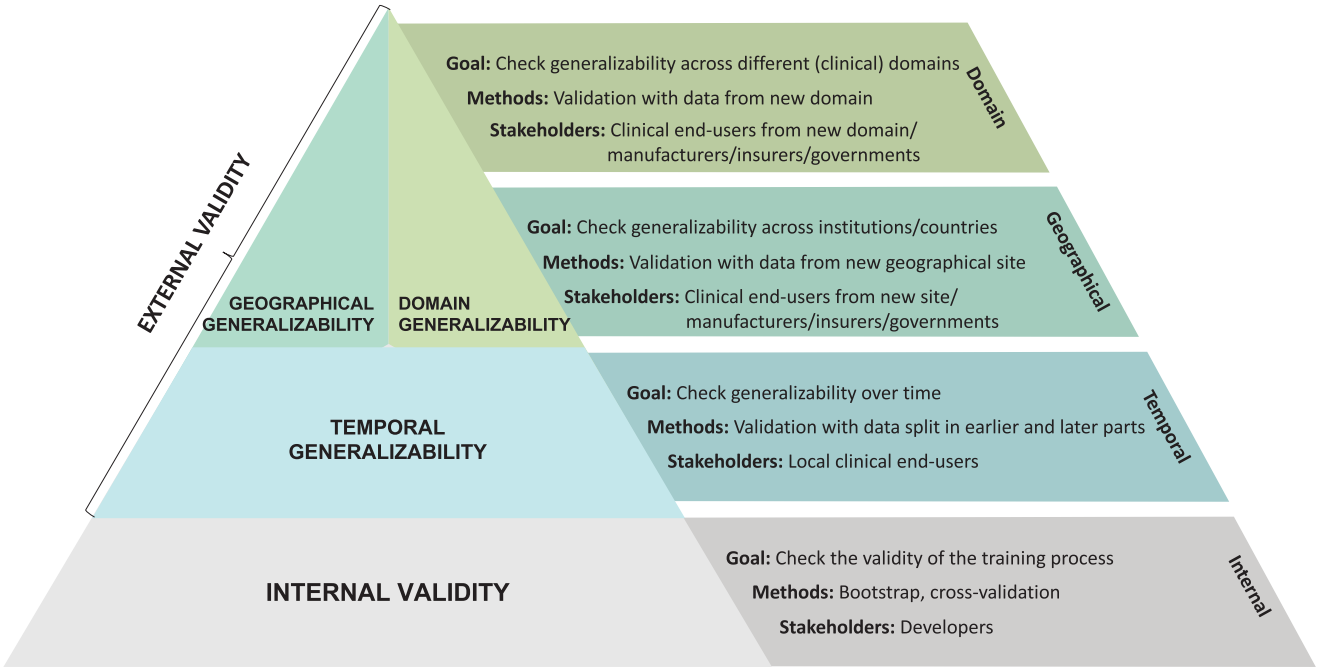

Fig. 1 Generalizability types. Schematic overview of the different types of generalizability with the validation's goals, methods, and stakeholders.

evidence for the general or widespread applicability of the prediction tool. When geographical generalizability is low, a global model that is valid for different places may not be tenable[13]. Instead, a local variant of the algorithm could be achieved through updating the global algorithm at each individual place[4].

Domain validation assesses the generalizability of an algorithm to a different clinical context[14,15]. This type of validation considers generalizability across medical background (e.g., 30-day mortality risk for emergency versus surgical patients), but also medical setting (e.g., fall prevention in nursing home versus hospital), and demographics (e.g., emergency admission risk for adult versus pediatric patients). For example, some COVID-19 prediction models were developed for related respiratory diagnoses[16]. In a large study on generalizability of prediction models, model performance was found to be better in 'closely related' than 'distantly related' validation cohorts, which underscores the relevance of domain generalizability[17]. Like geographical validation, domain validity is assessed by testing the algorithm on data collected from a new domain. Stakeholders of domain validity include clinical end-users from the new domain, manufacturers, insurers, and governing bodies. If the algorithm does not generalize across domains, the underlying relationships may be truly different, warranting separate algorithms for each domain.

The overview presented in Fig. 1 may be used as a starting point by regulatory bodies, industry, and academia when formulating guidelines and requirements for the generalizability of a clinical predictive algorithm. Building on previous work[18–21], we argue that validation studies should be suited to the target context and the intended use of the clinical predictive algorithm. Always aiming for a specific type of generalizability may not be defensible for some predictive algorithms and their intended use[18,22].

During algorithm development and validation, researchers and developers should adhere to guidelines, specifically TRIPOD or its forthcoming variant, TRIPOD-AI[20,23]. They should report on the algorithm's capacity to generalize and provide a justification for their chosen validation strategy by relating it to their intended operational period, (clinical) population, and environment. Moreover, they should add a disclaimer about the type of generalizability and intended use of their algorithms. If generalizability is limited this ought to be acknowledged alongside other implementation risks. For example, only internal or temporal validation was performed, or poor generalizability was found across places or clinical contexts. Researchers and developers should also report when the algorithm's scope limits the necessary validation steps. For example, domain validation may not be attempted when a predictive algorithm cannot be used (or has very limited use) outside of its domain (e.g., a prostate biopsy model).

In conclusion, we propose more precise specification for the desired and required type of generalizability for the implementation of clinical predictive algorithms. The three generalizability types discussed here, comprising temporal, geographical, and domain generalizability, all serve a unique goal and specific application purpose. Hence, researchers, developers, journals, funding organizations, and regulatory bodies should ensure that their chosen generalizability claims on the algorithm's intended use align with the underlying evidence. Future research may assess the impact of different types of heterogeneity on generalizability and steps to improve generalizability for clinical predictive algorithms.

Received: 10 January 2023; Accepted: 28 April 2023;
Published online: 06 May 2023

REFERENCES

1. Wu, E. et al. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat. Med.* **27**, 582–584 (2021).
2. Kakarmath, S. et al. Best practices for authors of healthcare-related artificial intelligence manuscripts. *npj Digital Med.* **3**, 134 (2020).
3. Steyerberg, E. W. & Vergouwe, Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur. Heart J.* **35**, 1925–1931 (2014).

4. Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
5. Vickers, A. J., Van Calster, B. & Steyerberg, E. W. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* **352**, i6 (2016).
6. Harrell, F. Multivariable modeling strategies. In: *Regression Modeling Strategies.* Springer Series in Statistics. (Springer, Cham., 2015).
7. Steyerberg, E. W. *Clinical prediction models* (Springer Nature, 2009).
8. Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap* (CRC press, 1994).
9. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digital Health* **2**, e489–e492 (2020).
10. Wan, B., Caffo, B. & Vedula, S. S. A unified framework on generalizability of clinical prediction models. *Front. Artif. Intell.* **5**, https://doi.org/10.3389/frai.2022.872720 (2022).
11. de Hond, A. A. H. et al. Predicting readmission or death after discharge from the ICU: external validation and retraining of a machine learning model. *Crit. Care Med.* **51**, 291–300 (2023).
12. Austin, P. C. et al. Geographic and temporal validity of prediction models: different approaches were useful to examine model performance. *J. Clin. Epidemiol.* **79**, 76–85 (2016).
13. Steyerberg, E. W., Nieboer, D., Debray, T. P. A. & van Houwelingen, H. C. Assessment of heterogeneity in an individual participant data meta-analysis of prediction models: an overview and illustration. *Stat. Med* **38**, 4290–4309 (2019).
14. Debray, T. P. et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J. Clin. Epidemiol.* **68**, 279–289 (2015).
15. Cowley, L. E., Farewell, D. M., Maguire, S. & Kemp, A. M. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagnostic Progn. Res.* **3**, 16 (2019).
16. Wynants, L. et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369**, m1328 (2020).
17. Gulati, G. et al. Generalizability of cardiovascular disease clinical prediction models: 158 independent external validations of 104 unique models. *Circ. Cardiovasc. Qual. Outcomes* **15**, e008487 (2022).
18. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F. & Celi, L. A. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* **2**, e489–e492 (2020).
19. Burns, M. L. & Kheterpal, S. Machine learning comes of age: local impact versus national generalizability. *Anesthesiology* **132**, 939–941 (2020).
20. de Hond, A. A. H. et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *npj Digital Med.* **5**, 2 (2022).
21. Sperrin, M., Riley, R. D., Collins, G. S. & Martin, G. P. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagnostic Progn. Res.* **6**, 24 (2022).
22. Van Calster, B., Steyerberg, E. W., Wynants, L. & van Smeden, M. There is no such thing as a validated prediction model. *BMC Med.* **21**, 70 (2023).
23. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur. Urol.* **67**, 1142–1151 (2015).

ACKNOWLEDGEMENTS

Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Award Number R01LM013362. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also supported by Research Foundation – Flanders (FWO) grant G097322N and Internal Funds KU Leuven grant C24M/20/064.

AUTHOR CONTRIBUTIONS

T.H.B. and A.A.H.d.H. conceived the idea. A.A.H.d.H. and V.B.S. performed the literature search and performed the analysis. A.A.H.d.H., V.B.S., I.M.J.K., B.v.C., E.W.S. and T.H.B. wrote the initial draft and approved the final manuscript. A.A.H.d.H. is the guarantor and accepts full responsibility for the work. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-023-00832-9.

Correspondence and requests for materials should be addressed to Anne A. H. de Hond.