ESC European Society of Cardiology

European Heart Journal (2023) **44**, 4831–4834
https://doi.org/10.1093/eurheartj/ehad727

SPECIAL ARTICLE
Digital health and innovation

Five critical quality criteria for artificial intelligence-based prediction models

Florien S. van Royen [1], Folkert W. Asselbergs [2,3], Fernando Alfonso [4], Panos Vardas[5], and Maarten van Smeden [6,7]*

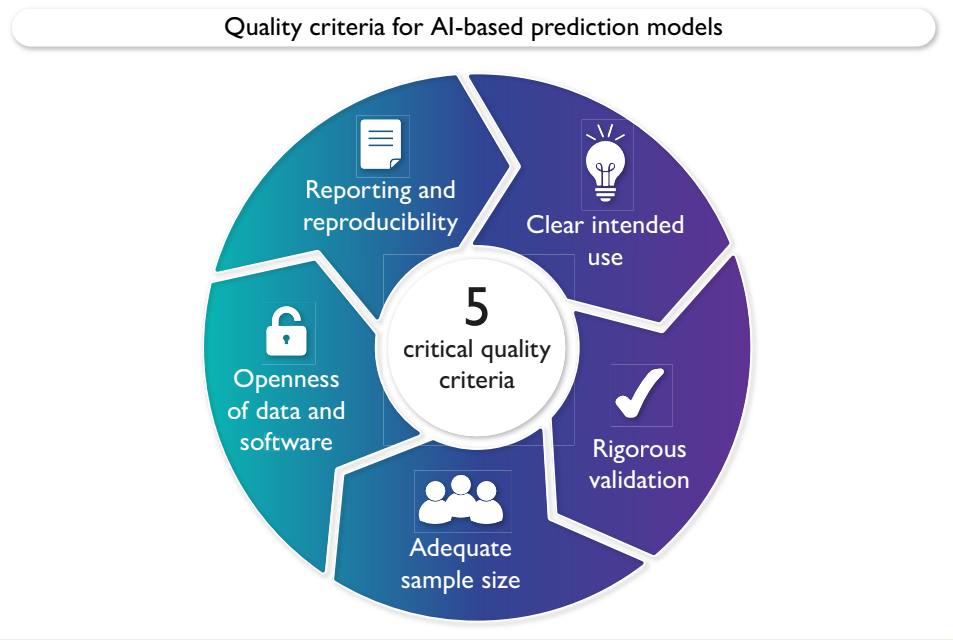
[1]Department of General Practice & Nursing Science, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands; [2]Department of Cardiology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, The Netherlands; [3]Health Data Research UK and Institute of Health Informatics, University College London, London, UK; [4]Department of Cardiology, Hospital Universitario de la Princesa, Universidad Autónoma de Madrid, IIS-IP. CIVER-CV, Madrid, Spain; [5]Biomedical Research Foundation Academy of Athens (BRFAA) and Hygeia Hospitals Group, Athens, Greece; [6]Department of Epidemiology & Health Economics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, Netherlands; and [7]Department of Data Science & Biostatistics, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG Utrecht, The Netherlands

Online publish-ahead-of-print 28 October 2023

Abstract

To raise the quality of clinical artificial intelligence (AI) prediction modelling studies in the cardiovascular health domain and thereby improve their impact and relevancy, the editors for digital health, innovation, and quality standards of the *European Heart Journal* propose five minimal quality criteria for AI-based prediction model development and validation studies: complete reporting, carefully defined intended use of the model, rigorous validation, large enough sample size, and openness of code and software.

Graphical Abstract



Five critical quality criteria for artificial intelligence (AI)-based prediction models.

* Corresponding author. Email: m.vansmeden@umcutrecht.nl

Keywords Artificial intelligence • Digital health • Prediction • Prognosis • Diagnosis

Introduction

As global cardiovascular disease burden is ever increasing, artificial intelligence (AI) holds great promise in reducing this burden through, among other ways, assisting in disease prevention by detection of at-risk individuals, offering more timely diagnoses and prognostication in patients, and reducing healthcare costs by automation of some of the tasks that were previously done by human experts.[1] Analytical AI techniques, such as neural networks and tree-based learning approaches, can handle large amounts of structured and unstructured forms of data (and their combination), and due to the many clinical data sources being available within cardiovascular medicine, such as physical examination results, laboratory results, imaging, electrocardiograms, and wearable devices, AI and machine learning techniques seem very suitable for use in cardiovascular health.[1]

In the cardiovascular health literature, analytical AI techniques are frequently used for the development of prediction models.[2] Despite the great potential of AI-based prediction models for application in the field of cardiovascular health, only few prediction models have so far shown their usefulness in clinical care.[3,4] To improve the chances of clinical implementation of AI-based prediction models and thus make impact on cardiovascular health, we must hold their development and validation to high scientific standards. In this paper we, as appointed editors for digital health, innovation, and quality standards of the *European Heart Journal*,[5] propose five minimal quality criteria that should be considered when developing a new AI-based prediction model. An extensive overview of critically reading and appraising cardiovascular disease prediction modelling research has been published recently in this journal.[6]

Quality criterion 1: complete reporting and reproducibility of results

Complete and transparent reporting is a key for reviewers and researchers to be able to fully appreciate and critically appraise the validity of model development methods and to evaluate the model's predictive performance. Furthermore, complete and transparent reporting improves replicability (similar results when re-developing and evaluating the model in different data sets) and reproducibility (similar results when repeating development in the original data), thereby improving credibility of the model. Systematic reviews have consistently shown that the reporting of prediction models, including those that are based on AI, is often poor.[7–9] Complete reporting should include the detailed description of all steps of the modelling process, including all data preparation steps, all model selection, tuning, recalibration, testing steps, and all results from internal and external validation procedures. To ensure all these elements are reported, relevant reporting guidelines should be used by authors, such as CODE-Electronic Healthcare Records (EHR) framework for structured electronic healthcare data and the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guideline for prediction model development, validation, and updating.[10–12] These reporting guidelines often come with a checklist that can be added as supplementary material to scientific manuscripts (e.g. see https://www.equator-network.org/reporting-guidelines/tripod-statement/). An update of the TRIPOD guidelines, specifically focused on AI-based prediction models, is expected soon.[13,14]

Quality criterion 2: clear intended clinical use of the AI-based model

The development of any AI-based prediction model should be motivated by a clearly defined clinical problem for which the AI prediction model could serve as a solution. The opportunities and possible pitfalls of a new AI-based model will only become evident if the intended use of the model, including where and how it should be positioned in the clinical workflow, is made explicit. Artificial intelligence-based prediction models can serve several purposes within cardiovascular health. For instance, the models can improve the diagnostic and prognostic clinical processes, by accurately predicting the presence of cardiovascular disease or predicting the progression of cardiovascular disease in a population of interest over a specific time frame.[15] Some well-known examples of prediction models for cardiovascular health are the Framingham risk score and the updated SCORE2.[16,17] The intended role of the AI-based prediction model in the clinical decision-making process, for instance in a prescriptive or assistive role, should be precisely defined to allow for early and careful consideration of the potential clinical consequences of using the model downstream in clinical care. A meeting with all relevant stakeholders, including physicians and patients, from the intended targeting in which the prediction model will be used in the future, can help identifying the potential impact, clinical requirements, and the potential for harm when implementing the model.[18]

Quality criterion 3: rigorous model validation

Model validation procedures ensure that the estimates of predictive performance of an AI-based prediction model, often summarized in terms of calibration and discrimination, are accurate and are estimated without over-optimism.[19–21] The estimates of performance obtained through *internal* validation techniques, such as cross-validation, reflect the expected performance when the model would be applied in (exactly) the same population—but in different individuals—than in which it was initially developed. The estimates of performance obtained through *external* validation techniques, for instance by applying the model in a separate dataset from a different region or hospital, reflect the performance in a different population from where the model was developed. These predictive performance estimates from external validation procedures may thus give an indication of the variation of performance of an AI-based model over time, place, and/or setting.[22] It should be noted that one external validation may not be sufficient to provide a complete picture of the heterogeneity of predictive performance, and therefore, all claims on model to be 'validated' should be viewed with some scepticism.[23] Good predictive performance also does not prove that the model will have a beneficial influence on medical decision-making when the model is used in a healthcare setting. For this, decision curve analysis, (early) health technology assessments, and impact studies (e.g. via randomized clinical trials) can generate valuable information on the clinical benefit and risks of an AI-based prediction model.[24,25]

Table 1 Summary of recommendations on artificial intelligence-based prediction models

	+	++	+++
Reporting and reproducibility	Following reporting guidelines (e.g. TRIPOD-AI)	Describing all steps of modelling and data processing	Providing guidance and open datasets to replicate/reproduce results
Clear intended use	Aim of the model stated clearly	Considering the downstream impact on clinical decision-making	Meeting with stakeholders about the potential barriers in prediction model use
Rigorous validation	Internal validation of the AI-based prediction models	Multiple internal and/or external validations of the AI-based prediction model	Rigorous evaluation of the variation of performance in multiple external and internal validations
Adequate sample size	A sample size for development that is substantially larger than needed for a regression-based prediction model	A posteriori sample size calculation (e.g. learning curves)	Sample size calculations for both model development and validation
Openness of data and software	Providing contact details for data and algorithm accessibility requests	Open software, including the code to apply the model in a new setting	Data and software publicly available

Quality criterion 4: sufficient sample size for AI model development and validation

Large enough sample sizes for both, the robust development and the accurate validation of the AI-based prediction model, are crucial. Calculators for regression-based prediction models to calculate the minimally required sample size may be useful starting points for AI-based prediction models.[26,27] However, due to the higher complexity of AI-based prediction models, the minimal required sample size may often be (much) larger, sometimes requiring data on multiple thousands of individuals, especially if the predicted outcome is rare (i.e. lower incidence or prevalence of the outcome to be predicted than 0.5 in the target population) and when the noise is high (i.e. low predictive effects of predictors and features). Currently, there are no calculators available that can be used to do a priori sample size calculations for the development of AI models. However, simulation studies and *a posteriori* approaches, such as a learning curve approach, may be used to justify the sample size.[22] For model validation studies, the minimally required sample size depends on the predictive performance criteria of the model and is not dependent on the modelling strategy. Therefore, sample size calculations can be performed a priori for validation studies and are the same for regression-based modelling as for AI modelling.[28]

Quality criterion 5: openness of data and software

Making the data and software—including the model code—publicly available is an important step in ensuring that readers and users can fully critically appraise the prediction model, perform tests (i.e. validations), and tailor the model to new settings. This will often increase the predictive performance of the model, the applicability, and clinical usefulness of the model and, eventually, improve model relevancy over time.[29] While we recognize the potential value methods from explainable AI (such as SHapley Additive exPlanations (SHAP) values) to give insights in what drives the predictions from an AI model (for some limitations, see[30,31]), it should not be viewed as a good replacement for sharing model code. Based on explainable AI output alone, a model cannot be externally validated.[30,31] Furthermore, while we recognize the important role of commercial parties in the field, which may have valid reasons to not fully share the model code (i.e. proprietary AI-based prediction models) and data used to develop and/or validate the AI-based prediction model, we warn against the tendency of researchers to not share code or data. Within the limitations given by commercial interests and privacy regulations, maximal openness of data and software should be strived for.[32] For a discussion on data sharing initiatives, we refer to earlier work in the *European Heart Journal*.[33]

Conclusion

This overview briefly touched upon five key quality criteria for authors, researchers, and readers of clinical AI prediction modelling studies in the field of cardiovascular health. A summary of the most important recommendations of this short viewpoint is provided in *Table 1* and *Graphical Abstract*. Complete reporting, carefully defined intended use of the model, rigorous validation, large enough sample sizes, and openness of code and software will increase the quality of clinical AI prediction studies and thereby the clinical impact and relevancy of their results.

Acknowledgements

F.W.A. is supported by UCL Hospitals NIHR Biomedical Research Centre, EU Horizon (AI4HF 101080430 and DataTools4Heart 101057849), and Dutch Research Council (MyDigiTwin 628.011.213).

Supplementary data

Supplementary data are not available at *European Heart Journal* online.

Declarations

Disclosure of Interest

All authors declare no disclosure of interest for this contribution.

Data Availability

No data were generated or analysed for this manuscript.

Funding

All authors declare no funding for this contribution.

References

1. Nakamura T, Sasano T. Artificial intelligence and cardiology: current status and perspective. *J Cardiol* 2022;**79**:326–33. https://doi.org/10.1016/j.jjcc.2021.11.017
2. Vardas PE, Asselbergs FW, van Smeden M, Friedman P. The year in cardiovascular medicine 2021: digital health and innovation. *Eur Heart J* 2022;**43**:271–9. https://doi.org/10.1093/eurheartj/ehab874
3. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, *et al.* Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ* 2016;**353**:i2416. https://doi.org/10.1136/bmj.i2416
4. Baart SJ, Dam V, Scheres LJJ, Damen JAAG, Spijker R, Schuit E, *et al.* Cardiovascular risk prediction models for women in the general population: a systematic review. *PLoS One* 2019;**14**:e0210329. https://doi.org/10.1371/journal.pone.0210329
5. Vardas P, Asselbergs F, van Smeden M. The new *European Heart Journal* digital health and innovations team. *Eur Heart J* 2021;**42**:1823–4. https://doi.org/10.1093/eurheartj/ehaa1087
6. Van Smeden M, Heinze G, Van Calster B, Asselbergs FW, Vardas PE, Bruining N, *et al.* Critical appraisal of artificial intelligence-based prediction models for cardiovascular disease. *Eur Heart J* 2022;**43**:2921–30. https://doi.org/10.1093/eurheartj/ehac238
7. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, *et al.* Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021;**375**:n2281. https://doi.org/10.1136/bmj.n2281
8. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;**368**:m689. https://doi.org/10.1136/bmj.m689
9. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, *et al.* Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *BMJ* 2020;**369**:m1328. https://doi.org/10.1136/bmj.m1328
10. 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. https://doi.org/10.7326/M14-0698
11. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015;**162**:55–63. https://doi.org/10.7326/M14-0697
12. Kotecha D, Asselbergs FW, Achenbach S, Anker SD, Atar D, Baigent C, *et al.* CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research. *Eur Heart J* 2022;**43**:3578–88. https://doi.org/10.1093/eurheartj/ehac426
13. Collins GS, Dhiman P, Andaur Navarro CL, 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**:e048008. https://doi.org/10.1136/bmjopen-2020-048008
14. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;**393**:1577–9. https://doi.org/10.1016/S0140-6736(19)30037-6
15. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. *J Clin Epidemiol* 2021;**132**:142–5. https://doi.org/10.1016/j.jclinepi.2021.01.009
16. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, *et al.* General cardiovascular risk profile for use in primary care: the Framingham heart study. *Circulation* 2008;**117**:743–53. https://doi.org/10.1161/CIRCULATIONAHA.107.699579
17. Hageman S, Pennells L, Ojeda F, Kaptoge S, Kuulasmaa K, de Vries T, *et al.* SCORE2 Risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J* 2021;**42**:2439–54. https://doi.org/10.1093/eurheartj/ehab309
18. Watson J, Hutyra CA, Clancy SM, Chandiramani A, Bedoya A, Ilangovan K, *et al.* Overcoming barriers to the adoption and implementation of predictive modeling and machine learning in clinical care: what can we learn from US academic medical centers? *JAMIA Open* 2020;**3**:167–72. https://doi.org/10.1093/jamiaopen/ooz046
19. 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. https://doi.org/10.1093/eurheartj/ehu207
20. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, *et al.* Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;**98**:691–8. https://doi.org/10.1136/heartjnl-2011-301247
21. Van Calster B, McLernon DJ, Van Smeden M, Wynants L, Steyerberg EW, Bossuyt P, *et al.* Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;**17**:230. https://doi.org/10.1186/s12916-019-1466-7
22. 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**:e007858. https://doi.org/10.1161/CIRCOUTCOMES.121.007858
23. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC Med* 2023;**21**:70. https://doi.org/10.1186/s12916-023-02779-w
24. 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. https://doi.org/10.1161/CIRCOUTCOMES.121.008487
25. Shah RU, Bress AP, Vickers AJ. Do prediction models do more harm than good? *Circ Cardiovasc Qual Outcomes* 2022;**15**:e008667. https://doi.org/10.1161/CIRCOUTCOMES.122.008667
26. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, *et al.* Sample size for binary logistic prediction models: beyond events per variable criteria. *Stat Methods Med Res* 2019;**28**:2455–74. https://doi.org/10.1177/0962280218784726
27. 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. https://doi.org/10.1136/bmj.m441
28. 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. https://doi.org/10.1002/sim.9025
29. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? *J Am Med Inform Assoc* 2019;**26**:1651–4. https://doi.org/10.1093/jamia/ocz130
30. McCoy LG, Brenna CTA, Chen SS, Vold K, Das S. Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *J Clin Epidemiol* 2022;**142**:252–7. https://doi.org/10.1016/j.jclinepi.2021.11.001
31. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3**: e745–50. doi: 10.1016/S2589-7500(21)00208-9
32. Van Calster B, Steyerberg EW, Collins GS. Artificial intelligence algorithms for medical prediction should be nonproprietary and readily available. *JAMA Intern Med* 2019;**179**:731. https://doi.org/10.1001/jamainternmed.2019.0597
33. Alfonso F; Editors' Network European Society of Cardiology Task Force; Editors' Network European Society of Cardiology Task Force. Data sharing. *Eur Heart J* 2017;**38**:1361–3. https://doi.org/10.1093/eurheartj/ehx206