Perspective

Predictive analytics in health care: how can we know it works?

Ben Van Calster,[1,2] Laure Wynants,[1] Dirk Timmerman,[1,3] Ewout W Steyerberg,[2] and Gary S Collins[4,5]

[1]Department of Development and Regeneration, KU Leuven, Leuven, Belgium, [2]Department of Biomedical Data Sciences, Leiden University Medical Center (LUMC), Leiden, The Netherlands, [3]Department of Obstetrics and Gynaecology, University Hospitals Leuven, Leuven, Belgium, [4]Centre for Statistics in Medicine, Nuffield, Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, UK, and [5]Oxford University Hospitals NHS Foundation Trust, Oxford, UK

Corresponding Author: Ben Van Calster, Department of Development and Regeneration, KU Leuven, Herestraat 49 box 805, 3000 Leuven, Belgium (ben.vancalster@kuleuven.be)

ABSTRACT

There is increasing awareness that the methodology and findings of research should be transparent. This includes studies using artificial intelligence to develop predictive algorithms that make individualized diagnostic or prognostic risk predictions. We argue that it is paramount to make the algorithm behind any prediction publicly available. This allows independent external validation, assessment of performance heterogeneity across settings and over time, and algorithm refinement or updating. Online calculators and apps may aid uptake if accompanied with sufficient information. For algorithms based on "black box" machine learning methods, software for algorithm implementation is a must. Hiding algorithms for commercial exploitation is unethical, because there is no possibility to assess whether algorithms work as advertised or to monitor when and how algorithms are updated. Journals and funders should demand maximal transparency for publications on predictive algorithms, and clinical guidelines should only recommend publicly available algorithms.

Key words: artificial intelligence, external validation, machine learning, model performance, predictive analytics

The current interest in predictive analytics for improving health care is reflected by a surge in long-term investment in developing new technologies using artificial intelligence and machine learning to forecast future events (possibly in real time) to improve the health of individuals. Predictive algorithms or clinical prediction models, as they have historically been called, help identify individuals at increased likelihood of disease for diagnosis and prognosis (see Supplementary Material Table S1 for a glossary of terms used in this manuscript).[1] In an era of personalized medicine, predictive algorithms are used to make clinical management decisions based on individual patient characteristics (rather than on population averages) and to counsel patients. The rate at which new algorithms are published shows no sign of abating, particularly with the increasing availability of Big Data, medical imaging, routinely collected electronic health records, and national registry data.[2–4] The scientific community is making efforts to improve data sharing, increase study registration beyond clinical trials, and make reporting transparent and comprehensive with full disclosure of study results.[5,6] We discuss the importance of transparency in the context of medical predictive analytics.

ALGORITHM PERFORMANCE IS NOT GUARANTEED: FULLY INDEPENDENT EXTERNAL VALIDATION IS KEY

Before recommending a predictive algorithm for clinical practice, it is important to know whether and for whom it works well. First,

predictions should discriminate between individuals with and without the disease (ie, higher predictions in those with the disease compared to those without the disease). Risk predictions should be also accurate (often referred to as calibrated).[7] Algorithm development may suffer from overfitting, which usually results in poorer discrimination and calibration when evaluated on new data.[8] Although the clinical literature tends to focus on discrimination, calibration is clearly crucial. Inaccurate risk predictions can lead to inappropriate decisions or expectations, even when discrimination is good.[7] Calibration has therefore been labeled the Achilles heel of prediction.[2]

In addition, there is often substantial heterogeneity between populations, as well as changes in populations over time.[9,10] For example, there may be differences between patients in academic hospitals compared with patients at regional hospitals, ethnicities, or past versus contemporary patients due to advances in patient care.[11–13] Recent work indicated that the half-life of clinical data relevance can be remarkably short.[14,15] Hence, algorithms are likely to perform differently across centers, settings, and time. On top of overfitting and heterogeneity between populations, operational heterogeneity can affect algorithm performance. Different hospitals may, for example, use different EHR software, imaging machines, or marker kits.[2,10,16] As a result, the clinical utility of predictive algorithms for decision-making may vary greatly. It is well established that "internal validation" of performance using, for example, a train–test split of available data is insufficient. Rather, algorithms should undergo "external validation" on a different data set.[17,18] Notably, algorithms developed using traditional study designs may not validate well when applied on electronic health record data.[4,19] It is important to stress 3 issues. First, external validation should be extensive: it should take place at various sites in contemporary cohorts of patients from the targeted population. Second, performance should be monitored over time.[11] Third, external validation by independent investigators is imperative.[20] It is a good evolution to include an external validation as part of the algorithm development study,[18] but one can imagine that algorithms with poor performance on a different data set may be less likely to get published in the first place. If performance in a specific setting is poor, an algorithm can be updated—specifically, its calibration.[1,7] To counter temporal changes in populations, continual updating strategies may help.[1] For example, QRISK2 models (www.qrisk.org) are updated regularly as new data are continually being collected.

POTENTIAL HURDLES FOR MAKING PREDICTIVE ALGORITHMS PUBLICLY AVAILABLE

To allow others to independently evaluate the predictive accuracy, it is important to describe in full detail how the algorithm was developed.[21] Algorithms should be available in a format that can readily be implemented by others. Not adhering to these principles severely limits the usefulness of the findings—surely a research waste.[22] An analogous situation would be an article describing the findings from a randomized clinical trial without actually reporting the intervention effect or how to implement the intervention.

Transparent and full reporting

The Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement, a reporting guideline for studies on predictive algorithms, recommends that the equation behind an algorithm is presented in the publication describing its development.[21] More explicitly, the mathematical formula of an algorithm should be available in full. This includes details such as which predictors are included, how they are coded (including ranges of any continuous predictors, units of measurement), and the values of the regression coefficients. Publications presenting new algorithms often fail to include key information such as specification of the baseline risk (namely, the intercept in logistic regression models for binary outcomes; the baseline hazard at 1 or more clinically relevant time points for time-to-event regression models).[23] Without this information, making predictions is not possible. Below, we expand on modern artificial intelligence methods that do not produce straightforward mathematical equations.

Online calculators and mobile apps

It has become customary to implement algorithms as online calculators or mobile apps. Then, we depend on the researchers' openness to provide clear and honest information about algorithm development and results of validation studies, with references to relevant publications. For example, FRAX predicts the 10-year probability of hip fracture and major osteoporotic fracture (www.sheffield.ac.uk/FRAX/). FRAX is a collection of algorithms (eg, 68 country-specific equations), which are both freely available via a website interface or commercially available via a desktop application. However, none of these algorithms has been published in full. The release notes indicate that the algorithms are continually revised, but do not offer detailed information. This lack of full disclosure prohibits independent evaluation.[24] In theory, we can try "reverse engineering" by reconstructing the equation based on risk estimates for a sample of patients (see Supplementary Material). However, such reverse engineering is not a realistic solution. The solution is to avoid hidden algorithms.

Online or mobile calculators allow the inclusion of algorithms into daily clinical routine, which is a positive evolution. However, it is impractical for large-scale independent validation studies, because information for every single patient has to be entered manually.

Machine learning algorithms

Machine learning methods, such as random forests or deep learning, are becoming increasingly popular to develop predictive algorithms.[3,25] The architecture of these algorithms is often too complex to fully disentangle and report the relation between a set of predictors and the outcome ("black box"). This is the commonly addressed problem when discussing transparency of predictive analytics based on machine learning.[26] We argue that algorithm availability is at least as important. A similar problem can affect regression-based algorithms that use complex spline functions to model continuous predictors. Software implementations are therefore imperative for validation purposes, in particular, because these algorithms have a higher risk of overfitting and instable performance.[8,17] Machine learning algorithms can be stored in computer files that may be transferred to other computers to allow validation studies. Recently, initiatives in this direction are being set up.[27,28]

Proprietary algorithms

Developers may choose not to disclose an algorithm, and to offer the algorithm on a fee-for-service basis.[16] For example, a biomarker-based algorithm to diagnose ovarian cancer has a cost of $897 per patient (http://vermillion.com/2436-2/). Assume we want to validate this algorithm in a center that has 20% malignancies in the target population. If we want to recruit at least 100 patients in each outcome group, following current recommendations for

Table 1. Summary of arguments in favor of making predictive algorithms fully available, hurdles for doing so, and reasons why developers choose to hide and sell algorithms

Why should predictive algorithms be fully and publicly available?	• Facilitate external validation and assessment of heterogeneity in performance • Facilitate uptake of algorithm by researchers and clinicians, avoid research waste • Facilitate updating for specific settings • For publicly funded research, this makes research results available to the community
Recommendations to maximize algorithm availability	• Report the full equation of a predictive algorithm, where possible (eg, regression-based models); this includes reporting of the intercept, or baseline hazard information for time-to-event regression models • When making an algorithm available online or via a mobile app, provide relevant and complete background information • For complex algorithms (eg, black-box machine learning), provide software to facilitate implementation and large-scale validation studies
Potential reasons why developers might choose to hide and sell algorithms	• Generate income for further research • More control over how people use an algorithm • Facilitate FDA approval or CE certification, because a commercial entity can be identified • To install a profitable business model

validation studies, the study needs at least 500 patients.[7] This implies a minimum cost of $448 500 in order to obtain useful information about whether this algorithm works in this particular center. It is important to emphasize this is just the cost required to judge whether the algorithm has any validity in this setting; there is no guarantee that it will be clinically useful.

Many predictive algorithms have been developed using financial support from public institutions. Then we believe that the results belong to the community and should be fully and publicly available. If this is the case, asking a small installation fee for an attractive and user-friendly calculator is defendable to cover software development and generate resources for maintenance and improvements. Such implementations facilitate uptake and inclusion into daily workflow.

Private companies may invest in the development of an algorithm that uses predictors for which the company offers measurement tools (eg, kits, biomarkers). In these instances, the return on investment should focus on the measurement tools, not on selling the algorithm. We argue that it is ethically unacceptable to have a business model that focuses on selling an algorithm.[29] However, such business models may facilitate Food and Drug Administration (FDA) approval or Conformité Européenne (CE) marking of predictive algorithms (eg, https://www.hcanews.com/news/predictive-patient-surveillance-system-receives-fda-clearance). It is important to realize that regulatory approval does not imply clinical validity or usefulness of a predictive algorithm in a specific clinical setting.[30]

THE IMPORTANCE OF ALGORITHM METADATA IN ORDER TO MAKE ALGORITHMS WORK

Although making algorithms fully and publicly available is imperative, the context of the algorithm is equally important. This extends the abovementioned issue of full and transparent reporting according to the TRIPOD guidelines. Reporting should provide full details of algorithm development practices. This includes—but is not limited to—the source of study data (e.g., retrospective EHR, randomized controlled trial data, or prospectively collected cohort data), the number and type of participating centers, the patient recruitment period, inclusion and exclusion criteria, clear definitions of predictors and the outcome, details on how variables were measured, detailed information on missing values and how these were handled,

and a full account of the modeling strategy (eg, predictor selection, handling of continuous variables, hyperparameter tuning). Unfortunately, studies reveal time and again that such metadata are poorly reported.[21,31] Even when authors develop an algorithm using sensible procedures (eg ,with low risk of overfitting), poor reporting will lead to poor understanding of the context, which may contribute to decreased performance on external validation. Initiatives such as the Observational Health Data Sciences and Informatics (OHDSI; http://ohdsi.org) focus on such contextual differences and aim to standardize procedures (eg, in terms of terminology, data formats, and definitions of variables) in order to lead to better and more applicable predictive algorithms.[27,32] In addition, when an algorithm is made available electronically, we recommend it include an indication of the extent to which the algorithm has been validated.

CONCLUSION

Predictive algorithms should be fully and publicly available to facilitate independent external validation across various settings (Table 1). For complex algorithms, alternative and innovative solutions are needed; a calculator is a minimal requirement, but downloadable software to batch process multiple records is more efficient. We believe that selling predictions from an undisclosed algorithm is unethical. This article does not touch on legal consequences of using predictive algorithms, where issues such as algorithm availability or black-box predictions cannot be easily ignored.[33] When journals consider manuscripts introducing a predictive algorithm, its availability should be a minimum requirement before acceptance. Clinical guideline documents should focus on publicly available algorithms that have been independently validated.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

FUNDING

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONTRIBUTIONS

Conception: BVC, LW, DT, EWS, GSC. Writing—original draft preparation: BVC. Writing—review and editing: BVC, LW, DT, EWS, GSC. All authors approved the submitted version and agreed to be accountable.

CONFLICT OF INTEREST STATEMENT

LW is a postdoctoral fellow of the Research Foundation – Flanders. GSC was supported by the NIHR Biomedical Research Centre, Oxford.

REFERENCES

1. Steyerberg EW. *Clinical Prediction Models*. New York: Springer; 2009.
2. Shah ND, Steyerberg EW, Kent DM. Big data and predictive analytics: recalibrating expectations. *JAMA* 2018; 320 (1): 27–8.
3. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; 319 (13): 1317–8.
4. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013; 20 (1): 117–21.
5. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JPA. Reproducible research practices and transparency across the biomedical literature. *PLoS Biol* 2016; 14 (1): e1002333.
6. Nosek BA, Alter G, Banks GC, et al. Promoting an open research culture. *Science* 2015; 348 (6242): 1422–5.
7. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 2016; 74: 167–76.
8. Lynch CJ, Liston C. New machine-learning technologies for computer-aided diagnosis. *Nat Med* 2018; 24 (9): 1304–5.
9. Riley RD, Ensor J, Snell KI, 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.
10. Ghassemi M, Naumann T, Schulam P, Beam AL, Ranganath R. Opportunities in machine learning for healthcare. arxiv: 1806.00388 [cs.LG]; 2018.
11. Davis SE, Lasko TA, Chen G, et al. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc* 2017; 24 (6): 1052–61.
12. Testa A, Kaijser J, Wynants L, et al. Strategies to diagnose ovarian cancer: new evidence from phase 3 of the multicenter international IOTA study. *Br J Cancer* 2014; 111 (4): 680–8.
13. Goff DC Jr, Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on practice guidelines. *J Am Coll Cardiol* 2014; 63 (25): 2935–59.
14. Chen JH, Alagappan M, Goldstein MK, et al. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int J Med Inform* 2017; 102: 71–9.
15. Levy-Fix G, Gorman SL, Sepulveda JL, Elhadad N. When to re-order laboratory tests? Learning laboratory test shelf-life. *J Biomed Inform* 2018; 85: 21–9.
16. He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019; 25 (1): 30–6.
17. van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *J Clin Epidemiol* 2016; 78: 83–9.
18. Nevin L; on behalf of the PLoS Medicine Editors. Advancing the beneficial use of machine learning in health care and medicine: toward a community understanding. *PLoS Med* 2018; 15 (11): e1002708.
19. Goldstein BA, Navar AM, Pencina MJ. Risk prediction with electronic health records. The importance of model validation and clinical context. *JAMA Cardiol* 2016; 1 (9): 976.
20. Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. *PLoS Med* 2013; 10 (2): e1001381.
21. Moons KGM, Altman DG, Reitsma JB, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162 (1): W1–73.
22. Glasziou P, Altman DG, Bossuyt P, et al. Reducing waste from incomplete and unusable reports of biomedical research. *Lancet* 2014; 383 (9913): 267–76.
23. Kleinrouweler CE, Cheong-See FM, Collins GS, et al. Prognostic models in obstetrics: available, but far from applicable. *Am J Obstet Gynecol* 2016; 214 (1): 79–90.
24. Collins GS, Michaëlsson K. Fracture risk assessment: state of the art, methodologically unsound, or poorly reported. *Curr Osteoporos Rep* 2012; 10 (3): 199–207.
25. Ohno-Machado L. Data science and artificial intelligence to improve clinical practice and research. *J Am Med Inform Assoc* 2018; 25 (10): 1273.
26. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018; 15 (11): e1002689.
27. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc* 2018; 25 (8): 969–75.
28. Wiegand T, Krishnamurthy R, Kuglitsch M, et al. WHO and ITU establish benchmarking process for artificial intelligence in health. *Lancet* 2019; 394 (10192): 9.
29. Char DS, Shah NH, Magnus D. Implementing machine learning in health care–addressing ethical challenges. *N Engl J Med* 2018; 378 (11): 981–3.
30. Park SH. Regulatory approval versus clinical validation of artificial intelligence diagnostic tools. *Radiology* 2018; 288 (3): 910–1.
31. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; 110: 12–22.
32. Hripcsak G, Ryan PB, Duke JD, et al. Characterizing treatment pathways at scale using the OHDSI network. *Proc Natl Acad Sci USA* 2016; 113 (27): 7329–36.
33. Black L, Knoppers BM, Avard D, et al. Legal liability and the uncertain nature of risk prediction: the case of breast cancer risk prediction models. *Public Health Genomics* 2012; 15 (6): 335–40.