BMC Research Notes

COMMENTARY Open Access

Reproducibility of prediction models in health services research

Lazaros Belbasis[1*] and Orestis A. Panagiotou[2,3,4]

Abstract

The field of health services research studies the health care system by examining outcomes relevant to patients and clinicians but also health economists and policy makers. Such outcomes often include health care spending, and utilization of care services. Building accurate prediction models using reproducible research practices for health services research is important for evidence-based decision making. Several systematic reviews have summarized prediction models for outcomes relevant to health services research, but these systematic reviews do not present a thorough assessment of reproducibility and research quality of the prediction modelling studies. In the present commentary, we discuss how recent advances in prediction modelling in other medical fields can be applied to health services research. We also describe the current status of prediction modelling in health services research, and we summarize available methodological guidance for the development, update, external validation and systematic appraisal of prediction models.

Keywords: Health services research, Machine learning, Open Science, Prediction modelling, Reproducibility, Research Methodology, Transparency

Introduction

Health services research is a multidisciplinary field that studies the health care system, including access to and delivery of care; the quality of the care provided to patients; the costs of care for patients, health systems, and payers; and ultimately the impact of care on health outcomes and well-being [1]. Data sources in health services research often differ from the traditional epidemiological investigations that prospectively or retrospectively collect data through active recruitment of participants based on a priori specified research questions [2]. Indeed, health services studies heavily rely on data that are routinely collected for purposes other than research, including health care billing claims, registry data, or electronic health records [1, 3]. Outcomes frequently examined by health services researchers include health care spending, and utilization of care services (e.g., hospital admission or readmission, admission to intensive care unit, length of hospitalization, or emergency department visit).

Making accurate predictions of these outcomes is crucial from the perspective of patients, clinicians, health economists, and policy makers. On the basis of prediction horizon, prediction models are classified into two categories: (a) diagnostic models (absence of a time horizon) and (b) prognostic models (presence of a time horizon). During the last decade, there has been an intensified discussion about the reproducibility of statistical methods for predicting outcomes in medicine, and more recently this discussion has expanded to prediction modelling using machine learning techniques [4]. Ensuring reproducible prediction models in health services research is critical for the deployment of these models in real-world settings to inform clinical and health policy decision-making.

*Correspondence: lazaros.belbasis@charite.de

[1] Meta-Research Innovation Center Berlin, QUEST Center, Berlin Institute of Health, Charité – Universitätsmedizin Berlin, Berlin, Germany
Full list of author information is available at the end of the article

In this commentary, we discuss recent advances in prediction modelling in fields such as clinical medicine that are relevant to ensuring reproducible models in health services research. While diagnostic modelling is common in health services studies (e.g., when developing algorithms for the accurate ascertainment of disease status from billing codes in administrative data), we focus here on prognostic models that predict a future outcome of interest over a time horizon, because these types of questions frequently concern health services problems. We present what is already known about prediction modelling in other medical fields, describe the current status of prediction modelling in health services research, and present recommendations and guidance to improve current research practices in this field.

Main text

Reproducibility and transparency in prediction modelling

Reproducibility, transparency, and openness are three interconnected concepts that are readily recognized as vital features of science [5–7]. Reproducibility is the ability of independent researchers to obtain the same (or similar) results when repeating an experiment or test, and it is considered a hallmark of high-quality science [8, 9]. Irreproducible research can occur because of practices applied in one or more steps involving study design, data quality, statistical analysis or study reporting [8]. Of direct relevance to prediction modelling (especially when machine learning methods are used) is computational reproducibility, which refers to the ability to repeat an analysis of a given dataset and obtain sufficiently similar results [10, 11]. It requires having available the complete analytical environment, including software, properly documented full source code, and the original data [10]. Ideally, the user and/or researcher should be able to inspect, modify and apply the code under modified parameter settings to reproduce the results and explore the robustness of the algorithm to the values of its parameters. In recent years, platforms designed for the development of software, such as GitHub, have been adopted by the scientific community as ways to distribute the code including many health services projects [10].

Transparency is another important component of high-quality research. Two major transparency measures are registration and pre-published protocols, which can reduce the selective reporting of prediction models. Although their importance in the context of randomized clinical trials is widely accepted and strongly promoted, their importance in prediction modelling research is not widely acknowledged [12]. Openness, a term including data and code sharing, is also a key indicator of high-quality research, but it remains an uncommon practice in prediction modelling research [12]. Promoting data and code sharing is expected to increase the number of external validation efforts and individual-participant data meta-analyses in prediction modelling for health services research. These processes can be enhanced by formalizing the data management and data sharing processes using the FAIR guiding principles for scientific data management and stewardship [13]. This document presents guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of data.

However, it should be acknowledged that data sharing in the context of health services research may be challenging. The main reason is that many routinely collected health data are provided to researchers for scientific purposes only upon approval by the entity that generates them (e.g., health insurer, health system etc.) under strict agreements to protect patient confidentiality and privacy [14]. Although these agreements often preclude further data sharing among researchers, data sharing could be facilitated through the development of large scientific consortia which have been successful in other research fields such as genetic epidemiology [15, 16].

Experience from prediction modeling in other research fields

Several large-scale systematic reviews of prediction models in clinical medicine have evaluated the quality of clinical prediction models and their potential to yield unbiased predictions. A summary of multiple risk-of-bias assessments examining more than 2000 models using PROBAST (i.e., a risk-of-bias assessment tool) showed that (a) two thirds of them have high risk of bias based on their statistical analysis, (b) one third of them had high risk of bias based on their outcome definition and ascertainment and (c) a quarter of them had high risk of bias based on how participants were selected [17]. Moreover, one of the largest systematic reviews of prediction models examined more than 400 models for outcome prediction in patients with chronic obstructive pulmonary disease [18]. The vast majority of the examined prediction models did not report the full model equation or any other form of model presentation. This is an important caveat of prediction models, because absence of any model presentation renders any effort to assess the reproducibility of a prediction model impossible and further diminishes the opportunity to deploy prediction models in routine settings, even if they have outstanding performance.

There have also been several assessments of machine learning models in areas outside health services research. Machine learning is a large family of statistical techniques with a rapidly increasing use in prediction modelling, especially in the field of health services research [1, 19]. Yet many prediction models based on machine learning methods have important limitations. For example,

their adherence to reporting guidelines is often suboptimal thereby reducing their potential to be reproduced and deployed in independent studies [20]. Additionally, a risk-of-bias assessment of multiple prediction models using supervised machine learning showed that almost 90% of the models were at high risk of bias [21]. Moreover, the handling of missing data is rarely reported, and when authors deal with missing data, they often poorly report the relevant methodological details [22]. These issues not only threaten the validity of statistical estimates but also make these models hard to reproduce. It is, therefore, important that health services researchers recognize these issues in advance and take proactive steps to ensure that prediction models addressing health services questions are not subject to similar limitations.

Prediction modelling in health services research

Numerous systematic reviews for prediction models have been published, and some of them focus on predicting outcomes relevant to health services research. For example, there are systematic reviews focused on prediction models for re-admission after an index hospitalization [23, 24], emergency hospital admission [25], length of hospital stay [26, 27], length of stay in the intensive care unit [28], and health care costs [29]. These systematic reviews summarize many prediction models, but their focus is on the data sources used, the predictors used and the model performance without providing a thorough assessment of reproducibility, transparency, and study quality.

Moreover, prediction models for outcomes relevant to health services research are often included in systematic reviews focusing on patients with a specific disease. For example, in a systematic review for patients with chronic obstructive pulmonary disease, 65 prediction models were identified for outcomes relevant to health services research (i.e., hospital admission, ICU admission, readmission after an index hospitalization, length of stay, and health care costs) [18]. However, there is a need for more systematic assessments of prediction models focusing exclusively on outcomes relevant to health services research. These systematic reviews could be used to draw important observations and recommendations to improve the development and validation of prediction models in this field. Of note, systematic reviews of prediction models should also adhere to the principles of open science to the extent possible. A starting point is pre-registration through relevant repositories or even journals that publish protocols of systematic reviews. For example, we recently published a protocol of a systematic review of multivariable models for prediction of health care spending using machine learning by following all the relevant frameworks and methodological guidance [30]. We hope that this research practice can become more prevalent in the near future.

Existing guidance for prediction models

A critical step in improving the reproducibility and research quality in prediction modelling for health services research is to systematically map the current research practices in this field. Through this process the issues contributing to irreproducibility and poor reporting of prediction models will be identified. However, to our knowledge, existing systematic reviews on prediction models for outcomes relevant to health services research have not performed a thorough assessment of prediction models.

Various frameworks are available for performing systematic reviews for prediction models, and we recommend that researchers follow them when conducting systematic reviews [31]. Also, the PRISMA statement is a general framework that was developed to guide any systematic review and meta-analysis in biomedical literature [32, 33]. To support the conduct of systematic reviews of prediction models, there is a validated search algorithm for prediction modelling studies in PubMed [34], and a guidance on how to construct a data extraction form [35]. Both these items can make the systematic review process more efficient, reproducible, and transparent. In addition, PROBAST, a risk-of-bias assessment tool for prediction modelling studies, can help contextualize biases arising from the selection of participants, the ascertainment of the outcome, the handling of predictors, and the statistical methods used for prediction [36, 37]. An extension of this tool (PROBAST-AI) for the assessment of prediction modelling studies using machine learning approaches is currently under development [38].

Researchers should also consider the life cycle of prediction modelling research, as it was previously described, before developing a new prediction model [39–42]. Based on the PROGRESS framework [39], the researchers should avoid developing new prediction models from scratch without ensuring that existing models are inadequate. Instead, when prediction models exist, they should aim to update them to improve their predictive performance and externally validate them to examine their generalizability in other populations. Moreover, before the deployment of prediction models in clinical practice or their use in decision-making, impact studies should be designed to assess their impact in real world settings [42].

The development, update and external validation of prediction models in health services research could be improved by following guidance that was developed during the last decade for clinical prediction models. Health services researchers building a prediction model should

follow the TRIPOD statement, which is a set of recommendations for the reporting of studies developing, validating or updating a prediction model and is endorsed by many journals [43, 44]. Although the TRIPOD statement was developed for traditional (parametric) statistical models, there is an ongoing process of developing the TRIPOD-AI statement, which will provide recommendations exclusively for machine learning models [38]. Also, there is additional guidance explaining how the prediction models should be presented [45].

Some additional guidance has been developed for prediction models using machine learning approaches. The MI-CLAIM checklist was developed to improve transparent reporting of machine learning algorithms in medicine, and it has similarities with TRIPOD statement [46]. Also, there is an additional framework on transparency, reproducibility, ethics, and effectiveness in machine learning applications for health [47]. Some standards for the computational reproducibility of machine learning models have been proposed, based on data, model and code publication, programming best practices and workflow automation [48, 49].

Outlook

Adhering to reproducible and transparent research practices when developing and employing a prediction model in health services research is important for the design of efficient health systems and health delivery programs, and the improvement in patients' outcomes. In this commentary, we summarize available frameworks and guidelines to develop, externally validate, update, and systematically review prediction models, and we discuss potential implications in health services research. These frameworks and approaches to reproducible prediction modelling that we discuss here require involvement from multiple stakeholders beyond individual researchers. Such stakeholders involve journal editors, peer-reviewers, funding bodies and universities, who can play a critical role in promoting, incentivizing and rewarding reproducible and transparent research practices.

Abbreviations

FAIR: Findability, accessibility, interoperability, and reuse; ICU: Intensive care unit; MI-CLAIM: Minimum information about clinical artificial intelligence modeling; PRISMA: Preferred reporting items for systematic reviews and meta-analyses; PROBAST: Prediction model risk of bias assessment tool; PROBAST-AI: Prediction model risk of bias assessment tool-artificial intelligence; TRIPOD: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis; TRIPOD-AI: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis-artificial intelligence.

Acknowledgements

Not applicable.

Author contributions

LB wrote the first draft of the manuscript and OAP critically commented on the first draft. Both authors read and approved the final manuscript.

Funding

No funding.

Availability of data and materials

Not applicable.

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

[1]Meta-Research Innovation Center Berlin, QUEST Center, Berlin Institute of Health, Charité – Universitätsmedizin Berlin, Berlin, Germany. [2]Center for Evidence Synthesis in Health, School of Public Health, Brown University, Providence, RI, USA. [3]Department of Health Services, Policy and Practice, School of Public Health, Brown University, Providence, RI, USA. [4]Department of Epidemiology, School of Public Health, Brown University, Providence, RI, USA.

Received: 31 March 2022 Accepted: 18 May 2022
Published online: 11 June 2022

References

1. Rose S. Intersections of machine learning and epidemiological methods for health services research. Int J Epidemiol. 2020;49(6):1763–70.
2. Belbasis L, Bellou V. Introduction to epidemiological studies. In: Evangelou E, editor. Genetic epidemiology: methods and protocols. New York: Humana Press; 2018. p. 1–6.
3. Panagiotou OA, Heller R. Inferential challenges for real-world evidence in the era of routinely collected health data. JAMA Oncol. 2021;7(11):1605–7.
4. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393(10181):1577–9.
5. McNutt M. Reproducibility. Science. 2014;343(6168):229.
6. Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, et al. Promoting transparency in social science research. Science. 2014;343(6166):30–1.
7. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348(6242):1422–5.
8. Resnik DB, Shamoo AE. Reproducibility and research integrity. Account Res. 2017;24(2):116–23.
9. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Sci Transl Med. 2016;8(341):341ps12.
10. Celi LA, Citi L, Ghassemi M, Pollard TJ. The PLOS ONE collection on machine learning in health and biomedicine: towards open code and open data. PLoS ONE. 2019;14(1):e0210232.
11. McDermott MBA, Wang S, Marinsek N, Ranganath R, Foschini L, Ghassemi M. Reproducibility in machine learning for health research: still a ways to go. Sci Transl Med. 2021;13(586):eabb1655.
12. Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KGM, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med. 2014;11(7):e1001671.
13. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3(1):160018.

14. Lane J, Schur C. Balancing access to health data and privacy: a review of the issues and approaches for the future. Health Serv Res. 2010;45(5p2):1456–67.
15. Austin MA, Hair MS, Fullerton SM. Research guidelines in the era of large-scale collaborations: an analysis of genome-wide association study consortia. Am J Epidemiol. 2012;175(9):962–9.
16. Budin-Ljøsne I, Isaeva J, Maria Knoppers B, Marie Tassé A, Shen H, McCarthy MI, et al. Data sharing in large research consortia: experiences and recommendations from ENGAGE. Eur J Hum Genet. 2014;22(3):317–21.
17. Jong Y, Ramspek CL, Zoccali C, Jager KJ, Dekker FW, Diepen M. Appraising prediction research: a guide and meta-review on bias and applicability assessment using the prediction model risk of bias assessment tool (PROBAST). Nephrology. 2021;26(12):939–47.
18. Bellou V, Belbasis L, Konstantinidis AK, Tzoulaki I, Evangelou E. Prognostic models for outcome prediction in patients with chronic obstructive pulmonary disease: systematic review and critical appraisal. BMJ. 2019;367:l5358.
19. Doupe P, Faghmous J, Basu S. Machine learning for health services researchers. Value Heal. 2019;22(7):808–15.
20. Dhiman P, Ma J, Navarro CA, Speich B, Bullock G, Damen JA, et al. Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. J Clin Epidemiol. 2021;138:60–72.
21. 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.
22. Nijman S, Leeuwenberg A, Beekers I, Verkouter I, Jacobs J, Bots M, et al. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. J Clin Epidemiol. 2022;142:218–29.
23. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission. JAMA. 2011;306(15):1688–98.
24. Artetxe A, Beristain A, Graña M. Predictive models for hospital readmission risk: a systematic review of methods. Comput Methods Progr Biomed. 2018;164:49–64.
25. Wallace E, Stuart E, Vaughan N, Bennett K, Fahey T, Smith SM. Risk prediction models to predict emergency hospital admission in community-dwelling adults: a systematic review. Med Care. 2014;52(8):751–65.
26. Lequertier V, Wang T, Fondrevelle J, Augusto V, Duclos A. Hospital length of stay prediction methods. Med Care. 2021;59(10):929–38.
27. Lu M, Sajobi T, Lucyk K, Lorenzetti D, Quan H. Systematic review of risk adjustment models of hospital length of stay (LOS). Med Care. 2015;53(4):355–65.
28. Verburg IWM, Atashi A, Eslami S, Holman R, Abu-Hanna A, de Jonge E, et al. Which models can i use to predict adult ICU Length of stay? A systematic review. Crit Care Med. 2017;45(2):e222-31.
29. Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised learning methods for predicting healthcare costs: systematic literature review and empirical evaluation. AMIA Annu Symp Proc. 2017;2017:1312–21.
30. Huang AW, Haslberger M, Coulibaly N, Galárraga O, Oganisian A, Belbasis L, et al. Multivariable prediction models for health care spending using machine learning: a protocol of a systematic review. Diagnostic Progn Res. 2022;6(1):4.
31. Debray TPA, Damen JAAG, Snell KIE, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460.
32. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160.
33. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
34. Geersing G-J, Bouwmeester W, Zuithoff P, Spijker R, Leeflang M, Moons K. Search filters for finding prognostic and diagnostic prediction studies in medline to enhance systematic reviews. PLoS ONE. 2012;7(2):e32844.
35. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: The CHARMS checklist. PLoS Med. 2014;11(10):e1001744.
36. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med. 2019;170(1):W1–33.
37. 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.
38. 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(7):e048008.
39. Steyerberg EW, Moons KGM, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
40. Royston P, Moons KGM, Altman DG, Vergouwe Y. Prognosis and prognostic research: developing a prognostic model. BMJ. 2009;338:b604.
41. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. BMJ. 2009;338:b605.
42. Moons KGM, Altman DG, Vergouwe Y, Royston P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. BMJ. 2009;338:b606.
43. 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(1):55–63.
44. 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(1):W1–73.
45. Bonnett LJ, Snell KIE, Collins GS, Riley RD. Guide to presenting clinical prediction models for use in clinical settings. BMJ. 2019;365:l1737.
46. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med. 2020;26(9):1320–4.
47. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ. 2020;368:l6927.
48. Heil BJ, Hoffman MM, Markowetz F, Lee S-I, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods. 2021;18(10):1132–5.
49. Panagiotou OA, Högg LH, Hricak H, Khleif SN, Levy MA, Magnus D, et al. Clinical application of computational methods in precision oncology. JAMA Oncol. 2020;6(8):1282–6.

Publisher's Note