Multivariable Prediction Models for Health Care Spending Using Machine Learning: a Protocol of a Systematic Review

Andrew W. Huang (andrew_huang2@brown.edu)  
Brown University  https://orcid.org/0000-0001-7552-7091

Martin Haslberger  
Berlin Institute of Health

Neto Coulibaly  
Brown University

Omar Galárraga  
Brown University

Arman Oganisian  
Brown University

Lazaros Belbasis  
Berlin Institute of Health

Orestis A. Panagiotou  
Brown University

Research Article

Keywords: health care spending, machine learning, prediction, systematic review

Posted Date: November 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1064149/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

With rising cost pressures on health care systems, machine-learning (ML) based algorithms are increasingly used to predict health care costs. Despite their potential advantages, the successful implementation of these methods could be undermined by biases introduced in the design, conduct, or analysis of studies seeking to develop and/or validate ML models. The utility of such models may also be negatively affected by poor reporting of these studies. In this systematic review, we aim to evaluate the reporting quality, methodological characteristics, and risk of bias of ML-based prediction models for individual-level health care spending.

Methods

We will systematically search PubMed and Embase to identify studies developing, updating, or validating ML-based models to predict an individual’s health care spending for any medical condition, over any time period, and in any setting. We will exclude prediction models of aggregate-level health care spending, models used to infer causality, models using radiomics or speech parameters, models of non-clinically validated predictors (e.g. genomics), and cost-effectiveness analyses without predicting individual-level health care spending. We will extract data based on the CHARMS checklist, previously published research, and relevant recommendations. We will assess the adherence of ML-based studies to the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) and examine the inclusion of transparency and reproducibility indicators (e.g. statements on data sharing). To assess the risk of bias, we will apply the Prediction model Risk Of Bias Assessment Tool (PROBAST). Findings will be stratified by study design, ML methods used, population characteristics, and medical field.

Discussion

Our systematic review will appraise the quality, reporting, and risk of bias of ML-based models for individualized health care cost prediction. This review will provide an overview of the available models and give insights into the strengths and limitations of using ML methods for the prediction of health spending.

Trial registration:

Not applicable.

Background
Total health spending per capita has increased exponentially in nearly every country with concerning implications for the sustainability of global health financing systems, and by extension healthcare delivery systems. (1, 2) Numerous approaches have been implemented for constraining the growth of health spending, including capitated payments and other value-based insurance designs. (3) However, these approaches often rely upon the efficient allocation of resources based on predictions of future health spending. A common example is the risk adjustment of health plan payments which aims to reallocate funding towards plans with enrolled beneficiaries that are predicted to have higher than average annualized cost of care. (4) In this context, the efficiency of the health financing system is dependent on prediction models that can accurately estimate individual health spending, and the development of such models is a key topic in the field of health services research.

Regression-based techniques have been the most commonly used methods to predict health spending. However, these parametric approaches typically rely on strong assumptions about the true data generating mechanism and have difficulty with sparse or missing data. (5, 6) With recent advances in computation, machine learning (ML) techniques are being applied more frequently in prediction models for health spending. This includes, for example, the classification of patients into groups based on predicted spending amounts ("high-cost" or "high spenders") or groups based on predicted changes in spending amounts over time ("cost bloomers"). (7, 8) Compared to prediction models developed with parametric techniques, ML algorithms, such as random decision forests, neural networks, and penalized regression, have theoretical advantages and are generally considered more efficient, because they rely on fewer assumptions and have the ability to learn adaptively from the data. (6, 9)

Despite their potential advantages, the successful implementation of ML methods to predict healthcare costs in routine settings could be undermined if their predictive performance is poor or leads to overly optimistic predictions. Various elements in the design, conduct, and analysis of ML models may introduce biases, including the lack of external validation to prevent overfitting, unrepresentative sampling, or unaccounted missing data. The utility of these models may be also adversely affected by poor or inadequate reporting of the studies in the increasing body of literature through which they are disseminated to potential users including payers, health systems, and also individuals.

Previous systematic reviews have suggested that the methodological and reporting quality of ML-based prediction models for clinical outcomes is suboptimal. (10–13) A previously published systematic review of supervised learning models for health care spending had a very narrow scope and did not include a thorough methodological assessment of the literature. (14) In our study, we aim to summarize all ML-based prediction models developed for the prediction of individual-level health care spending, assess their reporting, and appraise their risk of bias. Summarizing the findings of these studies and understanding how they are reported can provide important insights into the strengths and limitations of using ML methods for the prediction of health spending.

Methods
We designed this systematic review according to the Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) and relevant research guidance by Debray et al. (15, 16)

**Literature search**

We will systematically search PubMed and Embase from inception to September 16, 2021 to identify studies reporting the development, update, or external validation of an individualized patient prediction model using any ML methods to predict health care spending or changes in health care spending over any time period in any setting.

We will use the following search algorithm: ("machine learning" OR "statistical learning" OR "ensemble" OR "superlearner" OR "transfer learning" OR "classification and regression tree" OR "decision tree" OR "random forest" OR "naive bayes" OR "neural network*" OR "support vector machine" OR "gradient boosting machine" OR "K nearest neighbour" OR "clustering" OR "deep learning" OR "reinforced learning") AND ("high cost*" OR "medical cost*" OR "medical care cost*" OR "health care cost*" OR "healthcare cost*" OR "cost of care" OR "costs of care" OR "per capita cost*" OR "cost bloom" OR "patient spending*" OR "health care spending*" OR "healthcare spending*" OR "medical care spending*" OR "medical spending*" OR "high utilizer*" OR "high need*" OR "super utilizer*" OR "payment*" OR "expenditure*" OR "reimbursement*" OR "risk adjustment"). We will also perform a reference screening of all eligible articles to identify additional studies.

**Eligibility criteria**

Table 1 shows a detailed description of the Population, Intervention, Comparator, Outcome, Timing, and Setting (PICOTS) for this systematic review. To consider a study as eligible, we will follow the definition of a prediction modelling study as proposed by the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement. (17, 18) Accordingly, a study will be eligible if it reports on the development, update or external validation of a model/algorithm used for predicting an individual's health care spending as a function of two or more covariates. We will include prediction models that were developed, updated, or validated using ML techniques in patients with any medical condition and in any care setting or time period. We will include models examining binary, continuous, or categorical outcomes relevant to health care costs. We will consider as eligible any observational study (e.g., prospective or retrospective cohort studies and case-control studies), but will not include any randomized or observational studies designed to evaluate the impact of ML-based prediction models in health care spending.
Table 1
Key items for framing aim, search strategy, and study inclusion and exclusion criteria following the PICOTS framework

| Item        | Definition                                                                                                                                 |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| Population  | Patients with documented costs of health care services in any setting                                                                      |
| Intervention| Any prediction model designed to predict individual patient probabilities for incurring costs of health care services in any setting, or to predict probabilities for any changes in patient costs over time. |
| Comparator  | Not applicable                                                                                                                               |
| Outcomes    | Any cost-related outcome as reported by prediction models                                                                                   |
| Timing      | Any prediction horizon as reported by prediction models                                                                                     |
| Setting     | Any health care setting                                                                                                                     |

We will exclude articles (a) describing ML-based prediction models using ecological data to predict aggregate-level health care spending (e.g., county-level, or country-level); (b) building ML-based models with a primary goal of causal inference, which aim to estimate the change in one’s healthcare costs if a covariate of interest (e.g. insurance) changed from one level (e.g. commercial insurance) to a different level (e.g. public insurance); (c) applying a traditional statistical methods, such as linear regression, logistic regression or Cox regression for the prediction purposes; (d) presenting a systematic review of prediction models; (e) describing prediction models using radiomics or speech parameters; (f) models using biomarkers that are not clinically validated (e.g. genetic polymorphisms), and (g) performing cost-effectiveness analysis without predicting individual-level health care spending. Additionally, we will exclude conference abstracts, because they do not present a detailed description of their methods and their results, which would hinder a thorough methodological assessment. We will also exclude methodological articles that present a novel ML approach for prediction modelling without aiming at building an ML-prediction model for health care spending.

Data extraction

To facilitate the data extraction process, we will construct a standardized form by following the CHARMS checklist, previously published research, and relevant recommendations.(15, 19–22) From each eligible article, we will extract the population characteristics, geographic location, sample size (and number of events for binary outcomes), study design, predicted outcome and its definition, prediction horizon, and measures of model performance (discrimination, calibration, classification, overall performance). We will also extract the ML methods used in the final prediction model, whether the study included development, internal validation and/or external validation of the model, and whether any model presentation was available in the eligible studies. We will specifically evaluate whether the authors reported only apparent performance of a prediction model or examined overfitting by using internal validation. Also, we will examine whether a shrinkage method was applied in eligible studies and which method was used. We will consider that the authors adjusted for optimism sufficiently if they re-evaluated the performance of a model in internal validation and performed shrinkage as well. We will additionally record the data source
of predictors, whether there was any inclusion of routinely collected molecular predictors, and whether there were any criteria for manually including or excluding predictors from the final prediction model. Additionally, we will categorize external validation efforts into temporal and geographic validation. (23) For each eligible study, we also will examine whether the authors reported the presence of missing data on examined outcomes and/or predictors included in the prediction models; if so, we will record how missing data were treated. We will also extract information on how continuous predictors were handled and whether non-linear trends for continuous predictors were assessed.

**Risk of bias and reproducibility assessment**

We will appraise the presence of bias in the studies developing, updating or validating a prediction model by using Prediction model Risk Of Bias Assessment Tool (PROBAST), which is a risk of bias assessment tool designed for systematic reviews of diagnostic and prognostic prediction models. (24) It contains multiple questions categorized into four different domains: participants, predictors, outcome, and statistical analysis. Question responses are categorized as either “yes”, “probably yes”, “probably no”, “no”, or “no information”, depending on the characteristics of the study. If a domain contains at least one question signaled as “no” or “probably no”, it is considered high-risk. To be considered low-risk, a domain should contain all questions answered with “yes” or “probably yes”. Overall risk of bias is graded as low-risk when all domains are considered low-risk, and overall risk of bias is considered high-risk when at least one of the domains is considered high-risk.

Moreover, we will appraise the computational reproducibility of the eligible studies by following recently published reproducibility standards. (19, 20, 25) This assessment will be based on the availability of data, models, source codes and dependencies, and analysis plan. We will grade the reproducibility of eligible articles into three categories with varying degrees of rigor for computational reproducibility.

**Discussion**

As the frequency of applying ML-based prediction models in health economics and outcomes research increases, it is important to track and appraise the quality of studies that report their development in order to facilitate their successful implementation in the real world. To address this need, our systematic review will perform an exhaustive and comprehensive identification, summarization, and synthesis of multivariable models that use ML techniques to predict an individual's healthcare spending. In addition, we will assess the quality, reporting, and risk of bias of eligible ML-based models and potentially identify models that can be reliably used in the real world. Our findings will summarize the available models and give insights into the strengths and limitations of using ML methods to predict healthcare spending. Through thorough appraisal of the evidence base on ML models for healthcare spending, we will derive recommendations for improving the research practices in prediction modelling of health care spending.

**Abbreviations**

machine learning (ML)
Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS)

Population, Intervention, Comparator, Outcome, Timing, and Setting (PICOTS)

Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)

Prediction model Risk Of Bias Assessment Tool (PROBAST)

Declarations

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: Not applicable.

Competing interests: The authors declare that they have no competing interests.

Funding: Not applicable.

Authors’ contributions: Study design and methods were developed by AWH, LB, and OAP. Initial literature review and keyword search criteria were refined by AWH, MH, NC, LB, and OAP. All authors provided writing and critical revision of the manuscript. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

References

1. Dieleman JL, Sadat N, Chang AY, Fullman N, Abbafati C, Acharya P, et al. Trends in future health financing and coverage: future health spending and universal health coverage in 188 countries, 2016–40. Lancet [Internet]. 2018 May 5 [cited 2021 Oct 14];391(10132):1783–98. Available from: https://pubmed.ncbi.nlm.nih.gov/29678341/

2. Chang AY, Cowling K, Micah AE, Chapin A, Chen CS, Ikilezi G, et al. Past, present, and future of global health financing: A review of development assistance, government, out-of-pocket, and other private spending on health for 195 countries, 1995-2050. Lancet [Internet]. 2019 Jun 1 [cited 2021 Oct 14];393(10187):2233–60. Available from: https://pubmed.ncbi.nlm.nih.gov/31030984/

3. Conrad DA. The Theory of Value-Based Payment Incentives and Their Application to Health Care. Health Serv Res [Internet]. 2015 Dec 1 [cited 2021 Oct 14];50(Suppl 2):2057–89. Available from: https://pubmed.ncbi.nlm.nih.gov/26549041/

4. Layton TJ, McGuire TG, van Kleef RC. Deriving risk adjustment payment weights to maximize efficiency of health insurance markets. J Health Econ [Internet]. 2018 Sep 1 [cited 2021 Oct 14];61:93–110. Available from: /pmc/articles/PMC6471663/
5. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. Value Heal. 2015 Mar 1;18(2):137–40.

6. Rose S. Intersections of machine learning and epidemiological methods for health services research. Int J Epidemiol [Internet]. 2020 Dec 1 [cited 2021 Sep 9];49(6):1763–70. Available from: https://pubmed.ncbi.nlm.nih.gov/32236476/

7. Relles D, Ridgeway G, Carter G. Data mining and the implementation of a prospective payment system for inpatient rehabilitation. Heal Serv Outcomes Res Methodol. 2002;3(3–4):247–66.

8. Tamang S, Milstein A, Sørensen HT, Pedersen L, Mackey L, Betterton JR, et al. Predicting patient “cost blooms” in Denmark: A longitudinal population-based study. BMJ Open [Internet]. 2017 Jan 1 [cited 2021 Sep 9];7(1):e011580. Available from: https://bmjopen.bmj.com/content/7/1/e011580

9. Rose S. A Machine Learning Framework for Plan Payment Risk Adjustment. Health Serv Res [Internet]. 2016 Dec 1 [cited 2021 Sep 9];51(6):2358–74. Available from: https://pubmed.ncbi.nlm.nih.gov/26891974/

10. Groot OQ, Ogink PT, Lans A, Twining PK, Kapoor ND, DiGiovanni W, et al. Machine learning prediction models in orthopedic surgery: A systematic review in transparent reporting. J Orthop Res [Internet]. 2021 [cited 2021 Sep 9]; Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/jor.25036

11. 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 Oct 1;138:60–72.

12. Gelderblom ME, Stevens KYR, Houterman S, Weyers S, Schoot BC. Prediction models in gynaecology: Transparent reporting needed for clinical application. Eur J Obstet Gynecol Reprod Biol [Internet]. 2021 [cited 2021 Sep 9];0(0). Available from: http://www.ejog.org/article/S0301211521004139/fulltext

13. Wang W, Kiik M, Peek N, Curcín V, Marshall IJ, Rudd AG, et al. A systematic review of machine learning models for predicting outcomes of stroke with structured data. PLoS One [Internet]. 2020 Jun 1 [cited 2021 Sep 9];15(6). Available from: https://pubmed.ncbi.nlm.nih.gov/32530947/

14. 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 proceedings AMIA Symp [Internet]. 2017 [cited 2021 Sep 13];2017:1312–21. Available from: /pmc/articles/PMC5977561/

15. 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 [Internet]. 2014 [cited 2021 Sep 9];11(10). Available from: https://pubmed.ncbi.nlm.nih.gov/25314315/

16. 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 [Internet]. 2017 [cited 2021 Sep 9];356. Available from: https://pubmed.ncbi.nlm.nih.gov/28057641/
17. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Br J Surg [Internet]. 2015 Jan 27 [cited 2021 Sep 9];102(3):148–58. Available from: https://academic.oup.com/bjs/article/102/3/148/6136428

18. Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, et al. Poor reporting of multivariable prediction model studies: Towards a targeted implementation strategy of the TRIPOD statement. BMC Med [Internet]. 2018 Jul 19 [cited 2021 Sep 9];16(1):1–12. Available from: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-018-1099-2

19. Stevens LM, Mortazavi BJ, Deo RC, Curtis L, Kao DP. Recommendations for reporting machine learning analyses in clinical research. Circ Cardiovasc Qual Outcomes [Internet]. 2020 [cited 2021 Sep 9];13:6556. Available from: https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.120.006556

20. 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 [Internet]. 2020 Sep 9 [cited 2021 Oct 28];26(9):1320–4. Available from: https://www.nature.com/articles/s41591-020-1041-y

21. 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 [Internet]. 2019 [cited 2021 Sep 9];367. Available from: https://pubmed.ncbi.nlm.nih.gov/31585960/

22. Leisman DE, Harhay MO, Lederer DJ, Abramson M, Adjei AA, Bakker J, et al. Development and Reporting of Prediction Models: Guidance for Authors from Editors of Respiratory, Sleep, and Critical Care Journals. Crit Care Med [Internet]. 2020 [cited 2021 Sep 9];48(5):623–33. Available from: /pmc/articles/PMC7161722/

23. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: Validating a prognostic model. BMJ. 2009 May;338(7708):1432–5.

24. 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 [Internet]. 2019 Jan 1 [cited 2021 Sep 9];170(1):W1–33. Available from: https://pubmed.ncbi.nlm.nih.gov/30596876/

25. Heil BJ, Hoffman MM, Markowetz F, Lee SI, Greene CS, Hicks SC. Reproducibility standards for machine learning in the life sciences. Nat Methods [Internet]. 2021 Oct 1 [cited 2021 Oct 28];18(10):1132–5. Available from: https://pubmed.ncbi.nlm.nih.gov/34462593/

**Tables**

*Table 1. Key items for framing aim, search strategy, and study inclusion and exclusion criteria following the PICOTS framework*
| Item         | Definition                                                                 |
|--------------|-----------------------------------------------------------------------------|
| Population   | Patients with documented costs of health care services in any setting       |
| Intervention | Any prediction model designed to predict individual patient probabilities for incurring costs of health care services in any setting, or to predict probabilities for any changes in patient costs over time. |
| Comparator   | Not applicable                                                              |
| Outcomes     | Any cost-related outcome as reported by prediction models                    |
| Timing       | Any prediction horizon as reported by prediction models                      |
| Setting      | Any health care setting                                                     |