BMJ Open

Protocol for Development of a Reporting Guideline for Causal and Counterfactual Prediction Models

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059715
Article Type:	Protocol
Date Submitted by the Author:	01-Dec-2021
Complete List of Authors:	Xu, Jie; University of Florida Guo, Yi; University of Florida Wang, Fei; Weill Cornell Medicine at Cornell University Xu, Hua; The University of Texas Health Science Center at Houston Lucero, Robert; University of California - Los Angeles Bian, Jiang; University of Florida Prosperi, Mattia; University of Florida, Department of Epidemiology
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Information technology < BIOTECHNOLOGY & BIOINFORMATICS

Protocol for Development of a Reporting Guideline for Causal and Counterfactual Prediction Models

Jie Xu[1], Yi Guo[1], Fei Wang[2], Hua Xu[3], Robert Lucero[4], Jiang Bian[1], and Mattia Prosperi[5,*]

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA
[2]Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York, NY, USA
[3]School of Biomedical Informatics, University of Texas Health Science at Houston, Houston, TX, USA
[4]School of Nursing, University of California - Los Angeles, Los Angeles, CA, USA
[5]Department of Epidemiology, University of Florida, Gainesville, FL, USA
*e-mail: m.prosperi@ufl.edu

ABSTRACT

Introduction While there are protocols for reporting on observational studies (e.g., STROBE, RECORD), estimation of causal effects from both observational data and randomized experiments (e.g., AGREMA, CONSORT), and on prediction modelling (e.g., TRIPOD), none is purposely made for assessing the ability and reliability of models to predict counterfactuals for individuals upon one or more possible interventions, on the basis of given (or inferred) causal structures. This paper describes methods and processes that will be used to develop a reporting guideline for causal and counterfactual prediction models (tentative acronym: PRECOG).

Methods and Analysis PRECOG will be developed following published guidance from the EQUATOR network, and will comprise five stages. Stage 1 will be bi-weekly meetings of a working group with external advisors (active until stage 5). Stage 2 will comprise a scoping/systematic review of literature on counterfactual prediction modelling for biomedical sciences (registered in PROSPERO). In stage 3, a computer-based, real-time Delphi survey will be performed to consolidate the PRECOG checklist, involving experts in causal inference, statistics, machine learning, informatics and protocols/standards. Stage 4 will involve the write-up of the PRECOG guideline based on the results from the prior stages. Stage 5 will seek the peer-reviewed publication of the guideline, of the scoping/systematic review, and dissemination.

Ethics and Dissemination The authors follow the principles of the Declaration of Helsinki. The guideline development, starting with the working group of stage 1, will be initiated upon approval by an Institutional Review Board. The dissemination of PRECOG and its products, in addition to journal publications, will be done through conferences, websites, and social media. PRECOG can help researchers and policymakers to carry out and critically appraise causal and counterfactual prediction model studies. PRECOG will also be useful for designing interventions, and we anticipate further expansion of the guideline for specific areas, e.g., pharmaceutical interventions.

ARTICLE SUMMARY

Strengths and limitations of this study

- Several prediction models developed on observational data are often used for calculations of alternative scenarios and interventions (counterfactuals), such as changing behaviors, exposures or treatments, possibly resulting in harm because of underlying bias in the data
- Counterfactual prediction methods merge causal inference and statistical learning, thus providing useful frameworks for development of intervention/treatment optimization models
- The PRECOG guideline will fill a gap in reporting standard for counterfactual prediction modelling
- Even with rigorous study design, execution and reporting standard, causal claims made upon observational data analyses might be still mistaken by wrong assumptions or unmeasured, hidden bias

1 BACKGROUND

The increasing availability of large electronic health record data has led to an explosion in the development of prediction models –both traditional statistics and machine learning– for diagnostic, prognostic, and treatment optimization purposes. Despite of the availability of reporting guidelines, e.g., "transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" (TRIPOD) [1], the quality of many studies is low, as well as adherence to reporting standards, and there is often misinterpretation of the models' operating capabilities, with possible misuse and harm at the individual and/or population level [2,3]. One of the most common mistakes is to consider a prediction model readily usable for interventions on individuals, by changing certain variables with the intent to improve outcomes, i.e., calculating alternative scenarios or so-called counterfactuals. Since prediction models are often learnt from observational data, there is no guarantee that the strongest predictors are causing the outcome of interest and are not confounded, mediated by others, or actually concomitant causes of it. While such bias is not a problem for mere prediction in similar populations –since variables are not being changed with the intent to modify risk– it becomes problematic in new populations (even with high cross-validation results) [4] and when trying to optimize outcomes [5].

Thus, formal causal assessment is needed when developing prediction models on observational data to be used for alternative scenarios and interventions, i.e., counterfactual prediction models. The approaches from traditional statistics, computational science, and econometrics, including the potential outcomes framework [6], do-calculus and directed acyclic graphs (DAGs) [7], are often focused on estimating a population-level causal effect for a single interventional query (treatment or exposure), but in principle can be used to calculate individual treatment effects and counterfactuals. Machine learning has also been employed for counterfactual prediction [8,9]. Several off-the-shelf methodologies have been revisited, including deep learning [10–13], and random forests [14].

Given the rise in counterfactual prediction modelling studies, there is need for common grounds on model reporting, to improve on overall quality (albeit adhering to a protocol might be necessary, yet not sufficient condition to study quality), and specifically on transparency and reproducibility of results.

In the "Enhancing the quality and transparency of health research" (EQUATOR) network (https://www.equator-network.org/), there are guidelines specifically designed for reporting causal effects on randomized clinical trials (RCTs), e.g., "consolidated standards of reporting trials" (CONSORT) [15] and "a guideline for reporting mediation analyses of randomized trials and observational studies" (AGREMA) [16]. Reporting guidelines for observational studies also mention causal effects inference, e.g., "strengthening the reporting of observational studies in epidemiology Using Mendelian randomization" (STROBE-MR) [17], "reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology" (RECORD-PE) [18], and the "instrumental variable methods in comparative safety and effectiveness research" [19]. Outside of EQUATOR, thePatient-Centered Outcomes Research Institute (PCORI) (https://www.pcori.org/) provides "Standards for CausalInference Methods in Analyses of Data from Observational and Experimental Studies in Patient-Centered Outcomes Research" (https://tinyurl.com/4x55ad3t). Also, there are guidelines for estimating causal effects in pragmatic randomized trials [20].Overall, existing guidelines are not well fitted for causal and counterfactual prediction modelling, although a number of them contain elements that are directly related. Consequently, we aim to develop a new reporting guideline, which we tentatively name as "prediction and counterfactual modelling guidelines" (PRECOG). The focus of PRECOG is the development and validation of counterfactual prediction models, where one or more variables can be intervened upon, and will require declaration of causal assumptions as well validation of causal claims. PRECOG will also cover software implementation and interoperability. The primary use cases of PRECOG are expected to fall within biomedical sciences, but they could be applied to other fields such as psychology or economics.

2 METHODS/DESIGN

The authors follow the ethical research principles of the Declaration of Helsinki. The guideline development will be started upon approval (or exemption) by the Institutional Review Board (IRB) at University of Florida, Gainesville, USA.

PRECOG will be developed following published guidance from the EQUATOR network [21]. We will develop the guideline in five stages: (1) bi-weekly meeting of a working group; (2) scoping/systematic review of causal and counterfactual prediction modelling studies; (3) reporting checklist draft and Delphi exercise; (4) development of the final guideline; and (5) peer-review, publication and dissemination. These stages are drawn from prior, successful development studies, in primis the protocol used for the making of TRIPOD-AI and PROBAST-AI [22].

2.1 Stage 1: Working Group Setup and Meetings

The core working group is composed by the co-authors of this protocol description, who met bi-weekly (30-45 minutes) since September 13, 2021 to discuss the development of the protocol itself and the foreseen PRECOG reporting guideline.

After the public posting of the protocol description, and IRB approvals, the working group will be expanded with external advisors with expertise in biomedical informatics, (bio)statistics, causal inference, computer science, epidemiology, health economics, health outcome research, standards, and related areas. Each member of the core working group will identify one or more suitable external advisors, who will be invited to participate in the meeting and prompted to suggest further advisors. The list of advisors will also be used for Stage 3 (Delphi exercise). The working group will make best efforts to assure diversity, variety in career stages, and multicultural representation. The extended working group will also meet bi-weekly, and each meeting will ideally be composed of 3-7 people, with at least one external advisor present (otherwise be rescheduled). The working group will work on: (a) review of existing EQUATOR/PCORI reporting guidelines; (b) evaluation of the results of the scoping/systematic review of counterfactual prediction modelling studies for biomedical sciences; (c) drafting of the initial reporting checklist for the Delphi survey; (d) review of the survey and development of the final guideline; (e) manuscript writing; and (f) submission of the products to peer-review, publication and dissemination.

2.2 Stage 2: Literature Review of Counterfactual Prediction Modelling Studies

The purpose of the literature review is twofold: (1) to build a knowledge base on study design, methodological approaches, use cases and reporting commonalities among causal inference and counterfactual prediction studies in biomedical sciences; and (2) to help development of reporting items for PRECOG. A subset of the working group members will concentrate on the review. After determining the overarching objective, search criteria and performing an initial screening, the team will decide if a scoping review will be preferred to a systematic review [23]. The planned reporting statement of choice is the "preferred reporting items for systematic reviews and meta-analyses" (PRISMA) [24], which includes also an extension for scoping reviews, and the working group will register the work in the "prospective register of systematic reviews" (PROSPERO) [25].

2.3 Stage 3: Delphi Exercise

We will conduct a Delphi survey to review and refine the items of the PRECOG reporting checklist. Delphi participants will be identified initially through the professional network of the core working group and of the external advisors, and further via literature search (including but not limited to the scoping/systematic review), social media screening, and snowballing by the active participants. As for the expanded working group composition, participants will be invited from diverse and multicultural backgrounds and different countries. Invitees will include academics at various career stages, researchers and investigators from non-profit and for-profit organizations, program officers from national/federal funding agencies, entrepreneurs, health care professionals, journal editors, policy makers, health care regulators, and end-users of predictive models. The working group will also discuss and agree on a suitable sample size for the Delphi survey.

We will employ computer-based, real-time Delphi, which offers some operational advantages with respect to traditional multi-round Delphi techniques [26]. The working group will develop an initial reporting checklist for PRECOG, based on the EQUATOR developing standard, existing related guidelines (e.g., TRIPOD, PCORI), and an anonymous online survey will be created where each checklist item can be evaluated in relation to its importance and relevance for the guideline, using a five-point Likert scale, and a free text box for comments. Also, at the end of the survey, another text box will allow more generic comments and propositions, e.g., new items to be added to the checklist. When a participant consents to participate and completes the survey for the first time, they receive a summary of all the responses to date, and a code to access the survey again within the next three weeks. Each participant can see the updated results within that time frame and make changes to their responses if they deem so. The survey is closed after the required sample size is reached, or a maximum of two months are passed from the first recorded response.

At the end of the Delphi survey, the working group will review the results and consolidate the checklist. Items will need to reach 80% agreement from the panel in order to be accepted (or omitted) in the development of the final guideline. Eighty Percent was chosen as an appropriate cut off based on work by Lynn [27], who suggested that when at least 10 experts are involved in consensus development, at least 80% of the experts must agree on an item to achieve content validity. Statements that do not meet the 80% agreement will be discussed during the bi-weekly meetings, and dropped if no consensus is reached by the extended working group.

2.4 Stage 4: Development of the Guideline and Related Products

Upon finalization of the reporting checklist from the Delphi exercise, the extended working group will develop the full PRECOG guidelines. The manuscript will be posted to a public pre-print website, e.g., bioRxiv or medRxiv, before submission to a peer-review journal, and possibly presented as abstract/poster in major international conferences, e.g., the annual conference of the American Medical Informatics Association (AMIA) or the Society for Epidemiology Research (SER). It is expected that the PRECOG initiative will produce at least the following papers:

- Guideline development protocol (this work);
- Scoping/systematic review or causal and counterfactual prediction models in biomedical sciences;
- PRECOG guideline.

2.5 Stage 5: Publication and Dissemination Plan

After being posted on preprint servers, the aforementioned manuscripts will be submitted to peer-reviewed international journals for final publication. The authors' list will be determined on the basis of effective individual contributions, following the "contributor roles taxonomy" (CRediT) (https://casrai.org/credit/), and might include additional contributors other than the working group members and external advisors.

The dissemination strategy will be discussed during the bi-weekly meetings. In addition to conferences and publications, it is likely that social media platforms such as Twitter will be leveraged to inform on the PRECOG availability and utility.

3 CONCLUSION

The number of causal inference and counterfactual prediction modelling studies, along with software development, is increasing rapidly. PRECOG can help researchers and policymakers to carry out and critically appraise these studies and tools, besides providing model developers with a transparent and reproducible framework, and liaising with model updating and evidence synthesis projects. PRECOG will also be useful for designing interventions, and we anticipate further expansion of the guidelines for specific areas, e.g., pharmaceutical interventions. The guideline will be periodically reviewed to ensure consistency with the EQUATOR standards and with best methodological, operational scientific, and ethical practices.

ETHICS STATEMENTS

Patient consent for publication

Not required.

CONTRIBUTORSHIP STATEMENT

JX wrote and submitted the protocol description. YG performed an initial literature review on reporting standards. FW and HX performed initial literature review on counterfactual prediction models. RL advised on protocol procedures and ethical review. JB and MP conceived the idea.

ACKNOWLEDGMENTS

This work has been in part supported by National Institutes of Health (NIH) - National Institute of Allergy and Infectious Diseases (NIAID) grants no. R01AI145552 and R01AI141810 (Dr. Prosperi), by National Institute on Aging (NIA) grants no. R33AG062884-03 (Dr. Lucero and Dr. Prosperi) and 5R21AG068717-02 (Dr. Bian and Dr. Guo), by National Cancer Institute (NCI) grants no. 5R01CA246418-02, 3R01CA246418-02S1, 1R21CA245858-01A1, 3R21CA245858-01A1S1,and 1R21CA253394-01A1 (Dr. Bian and Dr. Guo), and by the Centers for Disease Control and Prevention (CDC) grant no. U18DP006512 (Dr. Bian, Dr. Guo and Dr. Prosperi).

COMPETING INTERESTS

None declared.

References

1. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. J. Br. Surg. 102, 148–158 (2015).
2. Van Calster, B., Wynants, L., Riley, R. D., van Smeden, M. & Collins, G. S. Methodology over metrics: current scientific standards are a disservice to patients and society. J. Clin. Epidemiol. 138, 219–226, DOI: https://doi.org/10.1016/j.jclinepi.2021.05.018 (2021).

3. Collins, G. S., van Smeden, M. & Riley, R. D. Covid-19 prediction models should adhere to methodological and reporting standards. Eur. Respir. J.56, DOI: 10.1183/13993003.02643-2020 (2020). https://erj.ersjournals.com/content/56/3/2002643.full.pdf.

4. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. D.Dataset Shift in Machine Learning (The MITPress, 2009).

5. Prosperi, M.et al.Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nat. Mach.Intell. 2, 369–375 (2020).

6. Rubin, D. Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66, 688–701(1974).

7. Pearl, J., Glymour, M. & Jewell, N. Causal Inference in Statistics: A Primer (Wiley, 2016).

8. Curth, A., Svensson, D., Weatherall, J. & van der Schaar, M. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021).

9. McConnell, K. J. & Lindner, S. Estimating treatment effects with machine learning. Heal. Serv. Res. 54, 1273–1282, DOI:https://doi.org/10.1111/1475-6773.13212 (2019).

10. Louizos, C.et al. Causal effect inference with deep latent-variable models. In Advances in neural information processing systems, 6446–6456 (2017).

11. Alaa, A. M., Weisz, M. & Van Der Schaar, M. Deep counterfactual networks with propensity-dropout.arXiv preprintarXiv:1706.05966(2017).

12. Yoon, J., Jordon, J. & van der Schaar, M. GANITE: estimation of individualized treatment effects using generative adversarial nets. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April30 - May 3, 2018, Conference Track Proceedings (OpenReview.net, 2018).

13. Ghosh, S., Boucher, C., Bian, J. & Prosperi, M. Propensity score synthetic augmentation matching using generative adversarial networks (pssam-gan).Comput. Methods Programs Biomed. Updat.1, 100020 (2021).

14. Lu, M., Sadiq, S., Feaster, D. J. & Ishwaran, H. Estimating individual treatment effect in observational data using random forest methods. J. Comput. Graph. Stat. 27, 209–219, DOI: 10.1080/10618600.2017.1356325 (2018). PMID: 29706752.

15. Schulz, K. F., Altman, D. G. & Moher, D. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 340, DOI: 10.1136/bmj.c332 (2010). https://www.bmj.com/content.

16. Lee, H.et al.A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies: TheAGReMA Statement. JAMA 326, 1045–1056, DOI: 10.1001/jama.2021.14075 (2021). https://jamanetwork.com/journals/jama/articlepdf/2784353/jama_lee_2021_sc_210004_1631818986.61722.pdf.

17. Skrivankova, V. W.et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. JAMA 326, 1614–1621, DOI: 10.1001/jama.2021.18236 (2021). https://jamanetwork.com/journals/jama/articlepdf/2785494/jama_skrivankova_2021_sc_210005_1635192360.12205.pdf.

18. Langan, S. M.et al.The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (record-pe).BMJ 363, DOI: 10.1136/bmj.k3532 (2018). https://www.bmj.com/content/363/bmj.k3532.full.pdf.

19. Brookhart, M. A., Rassen, J. A. & Schneeweiss, S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol. Drug Saf. 19, 537–554, DOI: https://doi.org/10.1002/pds.1908 (2010). https://onlinelibrary.wiley.com/doi/pdf/10.1002/pds.1908.

20. Murray, E. J., Swanson, S. A. & Hernán, M. A. Guidelines for estimating causal effects in pragmatic randomized trials (2019). 1911.06030.

21. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. PLoS medicine 7, e1000217 (2010).

22. Collins, G. S.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 11, DOI: 10.1136/bmjopen-2020-048008 (2021). https://bmjopen.bmj.com/content/11/7/e048008.full.pdf.

23. Munn, Z.et al.Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med. Res. Methodol. 18, 143, DOI: 10.1186/s12874-018-0611-x (2018).

24. Page, M. J.et al.The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, DOI:10.1136/bmj.n71 (2021). https://www.bmj.com/content/372/bmj.n71.full.pdf.

25. Booth, A.et al.The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst. Rev. 1,2, DOI: 10.1186/2046-4053-1-2 (2012).

26. Gnatzy, T., Warth, J., von der Gracht, H. & Darkow, I.-L. Validating an innovative real-time delphi approach - a methodological comparison between real-time and conventional delphi studies. Technol. Forecast. Soc. Chang. 78,1681–1694, DOI: https://doi.org/10.1016/j.techfore.2011.04.006 (2011). The Delphi technique: Past, present, and future prospects.

27. Lynn, M. R. Determination and quantification of content validity. Nurs. research (1986).

FIGURE LEGENDS

Figure 1. Flowchart of the PREdiction and COunterfactual modelling Guidelines (PRECOG) development.

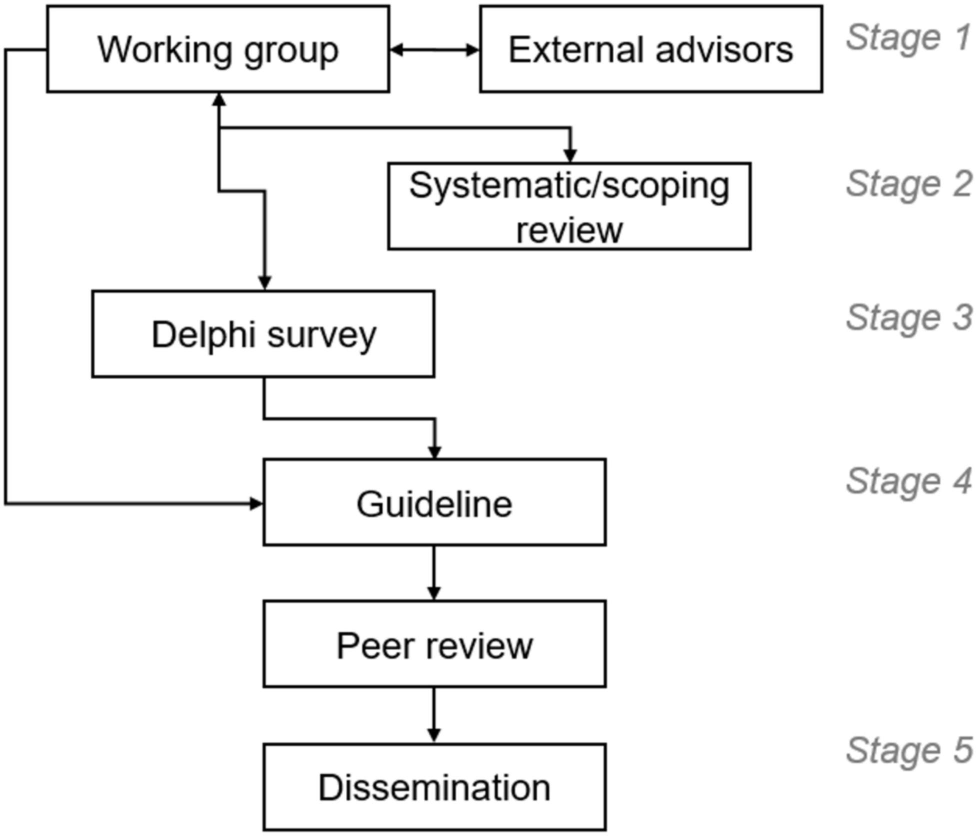

Figure 1. Flowchart of the PREdiction and COunterfactual modelling Guidelines (PRECOG) development.

109x93mm (600 x 600 DPI)

BMJ Open

Protocol for development of a reporting guideline for causal and counterfactual prediction models in biomedicine

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059715.R1
Article Type:	Protocol
Date Submitted by the Author:	06-Apr-2022
Complete List of Authors:	Xu, Jie; University of Florida Guo, Yi; University of Florida Wang, Fei; Weill Cornell Medicine at Cornell University Xu, Hua; The University of Texas Health Science Center at Houston Lucero, Robert; University of California - Los Angeles Bian, Jiang; University of Florida Prosperi, Mattia; University of Florida, Department of Epidemiology
<b>Primary Subject Heading</b>:	Health informatics
Secondary Subject Heading:	Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Information technology < BIOTECHNOLOGY & BIOINFORMATICS

Medical publishing and peer review

Protocol

Protocol for development of a reporting guideline for causal and counterfactual prediction models in biomedicine

Jie Xu[1], Yi Guo[1], Fei Wang[2], Hua Xu[3], Robert Lucero[4], Jiang Bian[1], and Mattia Prosperi[5,*]

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA

[2]Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York, NY, USA

[3]School of Biomedical Informatics, University of Texas Health Science at Houston, Houston, TX, USA

[4]School of Nursing, University of California - Los Angeles, Los Angeles, CA, USA

[5]Department of Epidemiology, University of Florida, Gainesville, FL, USA

*e-mail: m.prosperi@ufl.edu

ABSTRACT

Introduction While there are guidelines for reporting on observational studies (e.g., STROBE, RECORD), estimation of causal effects from both observational data and randomized experiments (e.g., AGREMA, CONSORT, PATH), and on prediction modeling (e.g., TRIPOD), none is purposely made for deriving and validating models from observational data to predict counterfactuals for individuals upon one or more possible interventions, on the basis of given (or inferred) causal structures. This paper describes methods and processes that will be used to develop a reporting guideline for causal and counterfactual prediction models (tentative acronym: PRECOG).

Methods and analysis PRECOG will be developed following published guidance from the EQUATOR network and will comprise five stages. Stage 1 will be meetings of a working group every other week with rotating external advisors (active until stage 5). Stage 2 will comprise a systematic review of literature on counterfactual prediction modeling for biomedical sciences (registered in PROSPERO). In stage 3, a computer-based, real-time Delphi survey will be performed to consolidate the PRECOG checklist, involving experts in causal inference, epidemiology, statistics, machine learning, informatics, and protocols/standards. Stage 4 will involve the write-up of the PRECOG guideline based on the results from the prior stages. Stage 5 will seek the peer-reviewed publication of the guideline, the scoping/systematic review, and dissemination.

Ethics and dissemination The authors follow the principles of the Declaration of Helsinki. The study has been registered in EQUATOR, and approved by the University of Florida's Institutional Review Board (#202200495); informed consent forms will be provided to both the working groups and the Delphi survey participants. The dissemination of PRECOG and its products, in addition to journal publications, will be done through conferences, websites, and social media.

Strengths and limitations of this study

- There are no guidelines for the reporting of data-learnt prediction models that have the specific intent to calculate alternative scenarios (counterfactuals) and identify individualized effects of interventions
- PRECOG will fill a gap in reporting standards for counterfactual prediction modeling and will capitalize on the systematization and quality of the EQUATOR network
- PRECOG will be built upon diverse (clinical researchers, computer scientists, epidemiologists, statisticians) expertise consensus across multiple development stages
- Even with rigorous study design, execution, and reporting standard, causal claims made upon observational data analyses might be still mistaken by wrong assumptions or unmeasured, hidden bias

INTRODUCTION

The increasing availability of large electronic health record data has led to an explosion in the development of prediction models –both traditional statistics and machine learning– for diagnostic, prognostic, and treatment optimization purposes. Despite the availability of reporting guidelines, e.g., "transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" (TRIPOD),[1] the quality of many studies is low, as well as adherence to reporting standards, and there is often a misinterpretation of the models' operating capabilities, with possible misuse and harm at the individual and/or population level.[2,3] One of the most common mistakes[4,5] is to consider a prediction model readily usable for interventions on individuals, by changing certain variables with the intent to improve outcomes, i.e., calculating alternative scenarios or so-called counterfactuals. Since prediction models are often learned from observational data, there is no guarantee that the strongest predictors are causing the outcome of interest and are not confounded, mediated by others, or actually concomitant causes of it. While such bias is not a problem for mere prediction in similar populations –since variables are not being changed with the intent to modify risk– it becomes problematic in new, out-of-distribution populations (even when cross-validation performance is high)[6] and when trying to optimize outcomes.[7]

Thus, formal causal assessment is needed when developing prediction models on observational data to be used for alternative scenarios and interventions, i.e., counterfactual prediction models. The approaches from traditional statistics, computational science, and econometrics, including the potential outcomes framework,[8] do-calculus, and directed acyclic graphs (DAGs),[9] are often focused on estimating a population-level causal effect for a single interventional query (treatment or exposure) but can be used to calculate individualized treatment effects and counterfactuals.[10–15] Machine learning has also been employed for counterfactual prediction.[16,17] Several off-the-shelf methodologies have been revisited, including deep learning,[18–20] and random forests.[21]

Given the rise in counterfactual prediction modeling studies, there is a need for common grounds on model reporting, to improve overall quality (albeit adhering to a protocol might be necessary, yet not sufficient condition to study quality), and specifically on transparency and reproducibility of results.

In the "Enhancing the quality and transparency of health research" (EQUATOR) network (https://www.equator-network.org/), there are guidelines specifically designed for reporting causal effects on randomized clinical trials (RCTs), e.g., "consolidated standards of reporting trials" (CONSORT)[22] and "a guideline for reporting mediation analyses of randomized trials and observational studies" (AGREMA).[23] Reporting guidelines for observational studies also mention causal effects inference, e.g., "strengthening the reporting of observational studies in epidemiology Using Mendelian randomization" (STROBE-MR),[24] "reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology" (RECORD-PE),[25] and the "instrumental variable methods in comparative safety and effectiveness research".[26] Outside of EQUATOR, the Patient-Centered Outcomes Research Institute (PCORI) (https://www.pcori.org/) provides "Standards for Causal Inference Methods in Analyses of Data from Observational and Experimental Studies in Patient-Centered Outcomes Research".[27] Also, there are guidelines for estimating causal effects in pragmatic randomized trials.[28] Worth noting is the “predictive approaches to treatment effect heterogeneity” (PATH) statement,[29] which –albeit focused on RCTs– examines risk-modeling within treatment (to analyze treatment outcome heterogeneity) as well as effect-modeling across treatment arms (to decide on treatment assignment). PATH provides guidance for specific multivariable regression configurations and warns against more “aggressive” approaches (e.g., machine learning models with many degrees of freedom) that could bring overfitting. Overall, existing guidelines are not well fitted for causal and counterfactual prediction modeling for observational biomedical data (or a mixture of RCTs and observational), although a number of them contain elements that are directly related.

Consequently, we aim to develop a new reporting guideline, which we tentatively name as PRECOG –acronym for "prediction of counterfactuals guideline". The primary focus of PRECOG is to provide guidance on how to report causal assumptions as well as evaluate derivation/validation of models that provide predictions of individualized treatment/intervention effects in the form of potential outcomes. On the one hand, the development of these models can follow both risk- and effect-modeling approaches as in PATH, but is intended to be more general allowing any functional form and data generation process. On the other hand, the validation standard of these models falls within the TRIPOD scopes, but it also evaluates how they are suitable for optimization (e.g., treatment decision, risk reduction) in addition to diagnosis and prognosis, trusting on the counterfactuals backed up by the causal claims. PRECOG is also expected to provide guidance on software implementation and interoperability. Translationally, PRECOG might be useful for designing interventions. By intervention, we mean that if a causal prediction model is used to evaluate alternative scenarios, then it can be used in a prospective way by public health officials or healthcare providers on individuals, reducing unfavorable health outcomes in the population overall. We anticipate further expansion of the guideline for specific areas, e.g., pharmaceutical interventions. As a quality evaluation instrument, PRECOG can also help researchers and policymakers to carry out and critically appraise causal and counterfactual prediction modeling studies. The primary use cases of PRECOG are expected to fall within biomedical sciences, but they could be applied to other fields such as psychology or economics.

METHODS AND ANALYSIS

PRECOG will be developed following published guidance from the EQUATOR network.[30] We will develop the guideline in five stages, as shown in **Figure 1**: (1) meeting of a working group every other week; (2) scoping/systematic review of causal and counterfactual prediction modeling studies; (3) reporting checklist draft and real-time Delphi exercise; (4) development of the final guideline; and (5) peer-review, publication, and dissemination. These stages are drawn from prior, successful development studies, in primis the protocol used for the making of TRIPOD-AI and PROBAST-AI.[31]

Stage 1: Working Group Setup and Meetings

The core working group is composed of the co-authors of this protocol description, who met every other week (30-45 minutes) since September 13, 2021, to discuss the development of the protocol itself, prepare documentation for the institutional review board (IRB), registration to EQUATOR, and eventually will carry out the PRECOG development after approvals and publication of the protocol description.

Then, the working group will be expanded with external advisors with expertise in biomedical informatics, (bio)statistics, causal inference, computer science, epidemiology, health economics, health outcome research, standards, and related areas. Each member of the core working group will identify one or more suitable external advisors, who will be invited to participate in the meeting and prompted to suggest further advisors, likely reaching 10-15 experts in total. The list of advisors will also be used for Stage 3 (real-time Delphi exercise). The expanded working group will make its best efforts to assure diversity, variety in career stages, geography, gender, race, and multicultural representation. The extended working group will also meet every other week, and each meeting will ideally be composed of 3-7 people, rotating participants, with at least one external advisor present (otherwise be rescheduled). The rotation and size limit of participants in a single meeting is built upon our prior experience with qualitative research, specifically focus groups, where compact size and diversified expertise aid to better reach data saturation.[32,33] The working group will work on: (a) review of existing EQUATOR/PCORI reporting guidelines related to prediction modeling and treatment effect estimation; (b) evaluation of published scoping reviews of counterfactual prediction modeling studies for biomedical sciences, and development of a new systematic review; (c) drafting of the initial reporting checklist for the Delphi survey; (d) review of the survey and development of the final guideline; (e) manuscript writing; and (f) submission of the products to peer-review, publication, and dissemination.

Stage 2: Literature Review of Counterfactual Prediction Modeling Studies

The purpose of the literature review is twofold: (1) to build a knowledge base on study design, methodological approaches, use cases, and reporting commonalities among causal inference and counterfactual prediction studies in biomedical sciences; and (2) to help the development of reporting items for PRECOG. A subset of the working group members will concentrate on the review. In 2021, Lin et al.[34] published a scoping review on causal methods for predictions under hypothetical interventions, screening nearly 5,000 papers and focusing on 13 key articles, including traditional statistical as well as machine learning modeling. Most works used marginal structural models and g-estimation. The authors concluded that “techniques for validating causal prediction models’ are still in their infancy.” Based on the results from the scoping review, and expanding the search strategy and the article sources, the team is going to move forward with a systematic review. The review will provide counts on methodology, review, and applied papers, but then will focus on works that include at least one observational data source and an application use case, further deepening the validation strategies. The planned reporting statement of choice is the "preferred reporting items for systematic reviews and meta-analyses" (PRISMA),[35] and the working group will register the work in the "prospective register of systematic reviews" (PROSPERO).[36]

As part of the review, we foresee discussing how to assess the potential risk of bias (which can lead to misuse and patients’ harm), and if current tools such as “prediction model risk of bias assessment tool” (PROBAST) are appropriate.[37]

Stage 3: Real-time Delphi Exercise

We will conduct a real-time Delphi survey[38] to review and refine the items of the PRECOG reporting checklist. Participants will be identified initially through the professional network of the core working group and of the external advisors, and further via literature search (including but not limited to the existing scoping review and the planned systematic review), social media screening, and snowballing by the active participants. As for the expanded working group composition, participants will be invited from diverse and multicultural backgrounds and different countries. Invitees will include academics at various career stages, researchers and investigators from non-profit and for-profit organizations, program officers from national/federal funding agencies, entrepreneurs, health care professionals, journal editors, policymakers, health care regulators, and end-users of predictive models. The participant selection will be based on area expertise grouping (computer science, biostatistics, biomedical informatics, statistics, epidemiology, standards, causal inference, ethics), used to determine the sample size (discussed below). We choose a computer-based, real-time Delphi,[38] since it offers some operational advantages with respect to conventional multi-round Delphi techniques, e.g., responder’s attrition.[39] In brief, real-time Delphi is a “roundless” exercise based on an online survey platform. Participants can access and modify their responses at any time during the survey timeframe, and they can view the survey summaries calculated among all responders. In this way, participants can see if/how their opinion is unpopular, and add further comments to support their cases.

The working group will develop an initial reporting checklist for PRECOG, based on the EQUATOR developing standard and existing related guidelines/statements. We anticipate that PRECOG will draw substantially from the reporting items of TRIPOD as well as the recommendations of PATH; however, we expect major differences rather than a simple merge. For instance, performance evaluation as recommended in TRIPOD should be modified to include specific metrics such as the Precision Estimation of Heterogeneous Effects (PEHE),[40] and emphasize out-of-distribution validation. Another important aspect is the causal assumptions. PATH relies on RCTs, where randomization supports the strong ignorability of treatment assignments, while PRECOG models might be exclusively built on observational data and a justification for causal claims will need to be provided.

An anonymous online survey will be created where each checklist item can be evaluated in relation to its importance and relevance for the guideline, using a five-point Likert scale, and a free text box for comments. Also, at the end of the survey, another text box will allow more generic comments and propositions, e.g., new items to be added to the checklist. When a participant consents to participate and completes the survey for the first time, they can view the summary of all responses to date and can access the survey again within the next six weeks. The survey is closed after the required sample size is reached, or a maximum of six weeks are passed from the last recorded first response.

There is no consensus on the sample size of a Delphi panel but a minimum number of 10-18 panel members per area of expertise has been recommended.[41] We will aim to reach a minimum sample size of 60 considering the aforementioned background expertise areas, compiling a list of 80-100 potential participants for the recruitment. At the end of the Delphi survey, the expanded working group will review the results and consolidate the checklist through a consensus meeting. The workgroup will also decide on the consensus rule. In general, for items ranked on a five-point Likert scale, the consensus rule is 80%,[42] but there can be differences in how adjacent items are grouped or weighted toward consensus.[43] For instance, Naughton et al.[44] quantified the Likert points from 1 (most important) to 5 (least important), and defined consensus for items scoring a median of 2.5 or less overall, when at least 80% of responders gave 1 to 3 points. More recent works proposed entropy-based consensus.[45]

Stage 4: Development of the Guideline and Related Products

Upon finalization of the reporting checklist from the Delphi exercise, the extended working group will develop the full PRECOG guidelines. The manuscript will be posted to a public pre-print website, e.g., bioRxiv or medRxiv, before submission to a peer-review journal, and possibly presented as an abstract/poster in major international conferences, e.g., the annual conference of the American Medical Informatics Association (AMIA) or the Society for Epidemiology Research (SER). It is expected that the PRECOG initiative will produce at least the following papers:

- Guideline development protocol (this work).
- Systematic review or causal and counterfactual prediction models in biomedical sciences.
- PRECOG guideline.

Stage 5: Publication and Dissemination Plan

After being posted on preprint servers, the aforementioned manuscripts will be submitted to peer-reviewed international journals for final publication. The authors' list will be determined based on effective individual contributions, following the "contributor roles taxonomy" (CRediT) (https://casrai.org/credit/), and might include additional contributors other than the working group members and external advisors.

The dissemination strategy will be discussed during the workgroup meetings. In addition to conferences and publications, it is likely that social media platforms such as Twitter will be leveraged to inform on the PRECOG availability and utility.

Ethics and dissemination

The authors follow the principles of the Declaration of Helsinki. The study has been registered in EQUATOR (https://tinyurl.com/2p88ucnb) and approved by University of Florida's IRB (protocol no. IRB202200495); informed consent forms will be provided to both the working groups and the Delphi survey participants. The dissemination of PRECOG and its products, in addition to journal publications, will be done through conferences, websites, and social media as described in the main text.

Acknowledgments We thank the TRIPOD co-authors Dr. G. Collins (U Oxford, UK) and Dr. K.G. Moons (UMC Utrecht, NL), and Dr. N. Peek (U Manchester, UK) for expressing their interest to join the PRECOG working groups.

Contributors JX wrote and submitted the protocol description. YG performed an initial literature review on reporting standards. FW and HX performed an initial literature review on counterfactual prediction models. RL advised on protocol procedures and ethical review. JB and MP conceived the idea.

Competing interests None declared.

Patient and public involvement This study does not include patients. However, the participants of the working groups –by definition– will be involved in the design of the Delphi survey, in its evaluation, and in the finalization of the PRECOG guideline (including authorship in papers). The participants of the Delphi survey can provide not only evaluation of items but suggest new ones and re-evaluate the items during the time when the survey is open.

References

1. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. Vol. 67, European Urology. 2015. p. 1142–51.

2. Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. J Clin Epidemiol. 2021 Oct;138:219–26.

3. Collins GS, van Smeden M, Riley RD. COVID-19 prediction models should adhere to methodological and reporting standards. Eur Respir J. 2020 Sep;56(3).

4. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. Vol. 32, CHANCE. 2019. p. 42–9.

5. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health. 2020 Dec;2(12):e677–80.

6. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset Shift in Machine Learning. MIT Press; 2008. 248 p.

7. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Vol. 2, Nature Machine Intelligence. 2020. p. 369–75.

8. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Vol. 66, Journal of Educational Psychology. 1974. p. 688–701.

9. Pearl J, Glymour M, Jewell NP. Causal Inference in Statistics: A Primer. John Wiley & Sons; 2016. 160 p.

10. Dorresteijn JAN, Visseren FLJ, Ridker PM, Wassink AMJ, Paynter NP, Steyerberg EW, et al. Estimating treatment

effects for individual patients based on the results of randomised clinical trials. Vol. 343, BMJ. 2011. p. d5888–d5888.

11. Nguyen VT, Rivière P, Ripoll P, Barnier J, Vuillemot R, Ferrand G, et al. Research response to coronavirus disease 2019 needed better coordination and collaboration: a living mapping of registered trials. Vol. 130, Journal of Clinical Epidemiology. 2021. p. 107–16.

12. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A. 2019 Mar 5;116(10):4156–65.

13. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. Vol. 30, Statistics in Medicine. 2011. p. 2867–80.

14. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012 Dec;68(4):1010–8.

15. Lamont A, Lyons MD, Jaki T, Stuart E, Feaster DJ, Tharmaratnam K, et al. Identification of predicted individual treatment effects in randomized clinical trials. Vol. 27, Statistical Methods in Medical Research. 2018. p. 142–57.

16. Brown K, Merrigan P, Royer J. Estimating Average Treatment Effects With Propensity Scores Estimated With Four Machine Learning Procedures: Simulation Results in High Dimensional Settings and With Time to Event Outcomes. SSRN Electronic Journal.

17. Hu L, Lin J-Y (Joyce), Sigel K, Kale M. Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. Ann Epidemiol. 2021 Oct;62:36–42.

18. Xiong M. Deep Learning for Causal Inference. Artificial Intelligence and Causal Inference. 2022. p. 151–208.

19. Ghosh S, Boucher C, Bian J, Prosperi M. Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN). Comput Methods Programs Biomed Update. 2021 Jul 16;1.

20. Ge Q, Huang X, Fang S, Guo S, Liu Y, Lin W, et al. Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection. Front Genet. 2020 Dec 11;11:585804.

21. Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. J Comput Graph Stat. 2018 Feb 1;27(1):209–19.

22. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2011 Oct 13;9(8):672–7.

23. Lee H, Cashin AG, Lamb SE, Hopewell S, Vansteelandt S, VanderWeele TJ, et al. A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies: The AGReMA Statement. JAMA. 2021 Sep 21;326(11):1045–56.

24. Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. JAMA. 2021 Oct 26;326(16):1614–21.

25. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ. 2018 Nov 14;363:k3532.

26. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf. 2010 Jun;19(6):537–54.

27. Seeger JD. Standards for Causal Inference Methods in Analyses of Data from Observational and Experimental Studies in Patient-Centered Outcomes Research. 2012.

28. Murray EJ, Swanson SA, Hernán MA. Guidelines for estimating causal effects in pragmatic randomized trials. arXiv preprint arXiv:1911.06030. 2019 Nov 14.

29. Baker SG. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020 Jun 2;172(11):775–6.

30. Moher D, Schulz KF, Simera I, Altman DG. Guidance for Developers of Health Research Reporting Guidelines.

Vol. 7, PLoS Medicine. 2010. p. e1000217.

31. 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 Jul 9;11(7):e048008.

32. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Soc Sci Med. 2022 Jan;292:114523.

33. Rich SN, Richards VL, Mavian CN, Switzer WM, Rife Magalis B, Poschman K, et al. Employing Molecular Phylodynamic Methods to Identify and Forecast HIV Transmission Clusters in Public Health Settings: A Qualitative Study. Viruses. 2020 Aug 22;12(9).

34. Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. Diagn Progn Res. 2021 Feb 4;5(1):3.

35. 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. Rev Esp Cardiol. 2021 Sep;74(9):790–9.

36. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Vol. 1, Systematic Reviews. 2012. Available from: http://dx.doi.org/10.1186/2046-4053-1-2

37. 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 Jan 1;170(1):W1–33.

38. Gordon T, Pease A. RT Delphi: An efficient, “round-less” almost real time Delphi method. Vol. 73, Technological Forecasting and Social Change. 2006. p. 321–33.

39. Hall DA, Smith H, Heffernan E, Fackrell K, for the Core Outcome Measures. Recruiting and retaining participants in e-Delphi surveys for core outcome set development: Evaluating the COMiT’ID study. PLoS One. 2018 Jul 30;13(7):e0201378.

40. Hill JL. Bayesian Nonparametric Modeling for Causal Inference. Vol. 20, Journal of Computational and Graphical Statistics. 2011. p. 217–40.

41. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. Vol. 42, Information & Management. 2004. p. 15–29.

42. Lynn MR. Determination and Quantification Of Content Validity. Vol. 35, Nursing Research. 1986.

43. Hsu C-C, Sandford BA. The Delphi Technique: Making Sense of Consensus. Practical Assessment, Research, and Evaluation. 2007;12(1):10.

44. Naughton B, Roberts L, Dopson S, Brindley D, Chapman S. Medicine authentication technology as a counterfeit medicine-detection tool: a Delphi method study to establish expert opinion on manual medicine authentication technology in secondary care. Vol. 7, BMJ Open. 2017. p. e013838.

45. Tastle WJ, Wierman MJ. Consensus and dissention: A measure of ordinal dispersion. Vol. 45, International Journal of Approximate Reasoning. 2007. p. 531–45.

FIGURE LEGENDS

Figure 1. Flowchart of the development of the reporting guideline for causal and counterfactual prediction models (PRECOG).

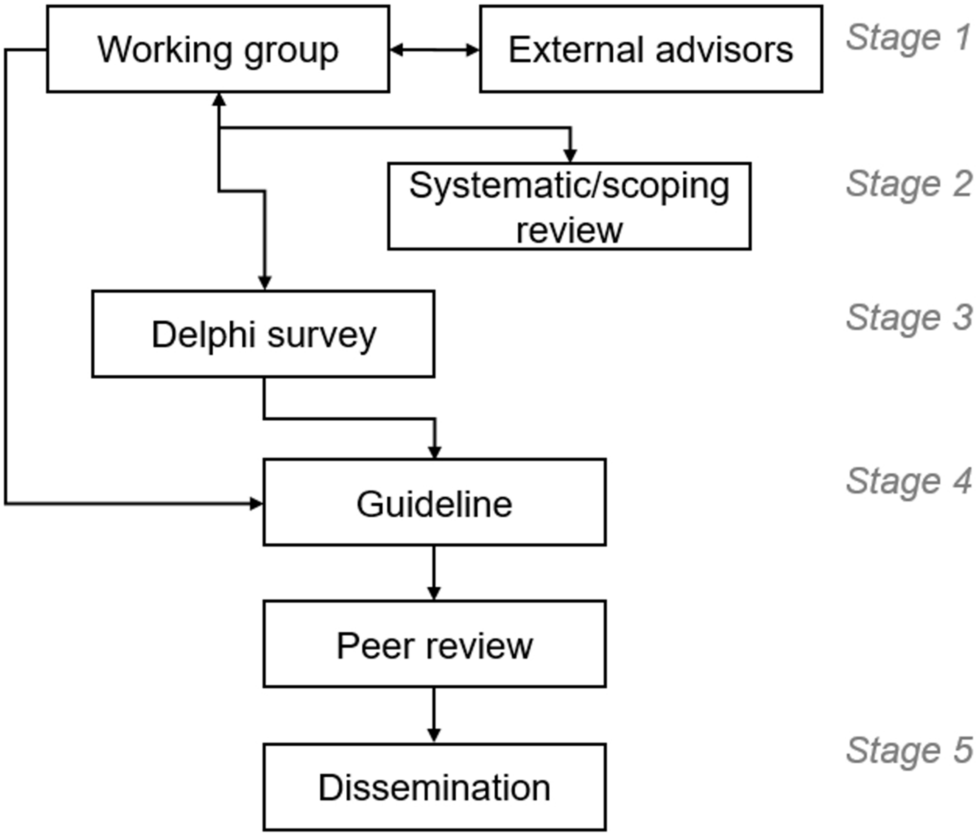

Figure 1. Flowchart of the development of the reporting guideline for causal and counterfactual prediction models (PRECOG).

109x93mm (765 x 765 DPI)

BMJ Open

Protocol for the development of a reporting guideline for causal and counterfactual prediction models in biomedicine

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-059715.R2
Article Type:	Protocol
Date Submitted by the Author:	20-May-2022
Complete List of Authors:	Xu, Jie; University of Florida Guo, Yi; University of Florida Wang, Fei; Weill Cornell Medicine at Cornell University Xu, Hua; The University of Texas Health Science Center at Houston Lucero, Robert; University of California - Los Angeles Bian, Jiang; University of Florida Prosperi, Mattia; University of Florida, Department of Epidemiology
<b>Primary Subject Heading</b>:	Health informatics
Secondary Subject Heading:	Health informatics
Keywords:	Health informatics < BIOTECHNOLOGY & BIOINFORMATICS, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Information technology < BIOTECHNOLOGY & BIOINFORMATICS

Protocol for the development of a reporting guideline for causal and counterfactual prediction models in biomedicine

Jie Xu[1], Yi Guo[1], Fei Wang[2], Hua Xu[3], Robert Lucero[4], Jiang Bian[1], and Mattia Prosperi[5,*]

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, USA
[2]Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York, NY, USA
[3]School of Biomedical Informatics, University of Texas Health Science at Houston, Houston, TX, USA
[4]School of Nursing, University of California - Los Angeles, Los Angeles, CA, USA
[5]Department of Epidemiology, University of Florida, Gainesville, FL, USA
*e-mail: m.prosperi@ufl.edu

ABSTRACT

Introduction While there are guidelines for reporting on observational studies (e.g., STROBE, RECORD), estimation of causal effects from both observational data and randomized experiments (e.g., AGREMA, CONSORT, PATH), and on prediction modeling (e.g., TRIPOD), none is purposely made for deriving and validating models from observational data to predict counterfactuals for individuals upon one or more possible interventions, on the basis of given (or inferred) causal structures. This paper describes methods and processes that will be used to develop a reporting guideline for causal and counterfactual prediction models (tentative acronym: PRECOG).

Methods and analysis PRECOG will be developed following published guidance from the EQUATOR network and will comprise five stages. Stage 1 will be meetings of a working group every other week with rotating external advisors (active until stage 5). Stage 2 will comprise a systematic review of literature on counterfactual prediction modeling for biomedical sciences (registered in PROSPERO). In stage 3, a computer-based, real-time Delphi survey will be performed to consolidate the PRECOG checklist, involving experts in causal inference, epidemiology, statistics, machine learning, informatics, and protocols/standards. Stage 4 will involve the write-up of the PRECOG guideline based on the results from the prior stages. Stage 5 will seek the peer-reviewed publication of the guideline, the scoping/systematic review, and dissemination.

Ethics and dissemination The study will follow the principles of the Declaration of Helsinki. The study has been registered in EQUATOR and approved by the University of Florida's Institutional Review Board (#202200495). Informed consent will be obtained from the working groups and the Delphi survey participants. The dissemination of PRECOG and its products will be done through journal publications, conferences, websites, and social media.

Strengths and limitations of this study

- There are no guidelines for the reporting of data-learnt prediction models that have the specific intent to calculate alternative scenarios (counterfactuals) and identify individualized effects of interventions.
- PRECOG will fill a gap in reporting standards for counterfactual prediction modeling and will capitalize on the systematization and quality of the EQUATOR network.
- PRECOG will be built upon diverse (clinical researchers, computer scientists, epidemiologists, statisticians) expertise consensus across multiple development stages.
- Even with rigorous study design, execution, and reporting standard, causal claims made upon observational data analyses might be still mistaken by wrong assumptions or unmeasured, hidden bias.

INTRODUCTION

The increasing availability of large electronic health record data has led to an explosion in the development of prediction models –both traditional statistics and machine learning– for diagnostic, prognostic, and treatment optimization purposes. Despite the availability of reporting guidelines, e.g., "transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" (TRIPOD),[1] the quality of many studies is low, as well as adherence to reporting standards, and there is often a misinterpretation of the models' operating capabilities, with possible misuse and harm at the individual and/or population level.[2,3] One of the most common mistakes[4,5] is to consider a prediction model readily usable for interventions on individuals, by changing certain variables with the intent to improve outcomes, i.e., calculating alternative scenarios or so-called counterfactuals. Since prediction models are often learned from observational data, there is no guarantee that the strongest predictors are causing the outcome of interest and are not confounded, mediated by others, or actually concomitant causes of it. While such bias is not a problem for mere prediction in similar populations –since variables are not being changed with the intent to modify risk– it becomes problematic in new, out-of-distribution populations (even when cross-validation performance is high)[6] and when trying to optimize outcomes.[7]

Thus, formal causal assessment is needed when developing prediction models on observational data to be used for alternative scenarios and interventions, i.e., counterfactual prediction models. The approaches from traditional statistics, computational science, and econometrics, including the potential outcomes framework,[8] do-calculus, and directed acyclic graphs (DAGs),[9] are often focused on estimating a population-level causal effect for a single interventional query (treatment or exposure) but can be used to calculate individualized treatment effects and counterfactuals.[10–15] Machine learning has also been employed for counterfactual prediction.[16,17] Several off-the-shelf methodologies have been revisited, including deep learning,[18–20] and random forests.[21]

Given the rise in counterfactual prediction modeling studies, there is a need for common grounds on model reporting, to improve overall quality (albeit adhering to a protocol might be necessary, yet not sufficient condition to study quality), and specifically on transparency and reproducibility of results.

In the "Enhancing the quality and transparency of health research" (EQUATOR) network (https://www.equator-network.org/), there are guidelines specifically designed for reporting causal effects on randomized clinical trials (RCTs), e.g., "consolidated standards of reporting trials" (CONSORT)[22] and "a guideline for reporting mediation analyses of randomized trials and observational studies" (AGREMA).[23] Reporting guidelines for observational studies also mention causal effects inference, e.g., "strengthening the reporting of observational studies in epidemiology Using Mendelian randomization" (STROBE-MR),[24] "reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology" (RECORD-PE),[25] and the "instrumental variable methods in comparative safety and effectiveness research".[26] Outside of EQUATOR, the Patient-Centered Outcomes Research Institute (PCORI) (https://www.pcori.org/) provides "Standards for Causal Inference Methods in Analyses of Data from Observational and Experimental Studies in Patient-Centered Outcomes Research".[27] Also, there are guidelines for estimating causal effects in pragmatic randomized trials.[28] Worth noting is the “predictive approaches to treatment effect heterogeneity” (PATH) statement,[29] which –albeit focused on RCTs– examines treatment effect heterogeneity by considering as effect modifier(s) either the risk or the covariates, with both strategies aimed at guiding treatment decisions. PATH provides guidance for specific multivariable regression configurations and warns against more “aggressive” approaches (e.g., machine learning models with many degrees of freedom) that could bring overfitting. Overall, existing guidelines are not well fitted for causal and counterfactual prediction modeling for observational biomedical data (or a mixture of RCTs and observational), although a number of them contain elements that are directly related.

Consequently, we aim to develop a new reporting guideline, which we tentatively name as PRECOG –acronym for "prediction of counterfactuals guideline". The primary focus of PRECOG is to provide guidance on how to report causal assumptions as well as evaluate derivation/validation of models –involving at least an observational data source– that provide predictions of individualized treatment/intervention effects in the form of potential outcomes. On the one hand, the development of these models can follow both risk- and effect-modeling approaches as in PATH, but it is intended to be more general, allowing any functional form and data generation process. On the other hand, the validation standard of these models falls within the TRIPOD scopes, but it also evaluates how they are suitable for optimization (e.g., treatment decision, risk reduction) in addition to diagnosis and prognosis, trusting on the counterfactuals backed up by the causal claims. PRECOG is also expected to provide guidance on software implementation and interoperability. As a quality evaluation instrument, PRECOG can help researchers (and general readers, peer reviewers, journal editors) as well as policymakers to carry out and critically appraise causal and counterfactual prediction modeling studies. We anticipate further expansion of the guideline for specific areas, e.g., pharmaceutical interventions. The primary use cases of PRECOG are expected to fall within biomedical sciences, but they could be applied to other fields such as psychology or economics.

METHODS AND ANALYSIS

PRECOG will be developed following published guidance from the EQUATOR network.[30] We will develop the guideline in five stages, as shown in **Figure 1**: (1) meeting of a working group every other week; (2) scoping/systematic review of causal and counterfactual prediction modeling studies; (3) reporting checklist draft and real-time Delphi exercise; (4) development of the final guideline; and (5) peer-review, publication, and dissemination. These stages are drawn from prior, successful development studies, in primis the protocol used for the making of TRIPOD-AI and PROBAST-AI.[31] The expected timeline for stages 1-4 is one year, using six-to-nine months for stages 1-2, and three-to-six months for stages 3-4.

Stage 1: Working group setup and meetings

The core working group is composed of the co-authors of this protocol description, who met every other week (30-45 minutes) since September 13, 2021, to discuss the development of the protocol itself, prepare documentation for the institutional review board (IRB), registration to EQUATOR, and eventually will carry out the PRECOG development after approvals and publication of the protocol description.

Then, the working group will be expanded with external advisors with expertise in biomedical informatics, (bio)statistics, causal inference, computer science, epidemiology, health economics, health outcome research, standards, and related areas. Each member of the core working group will identify one or more suitable external advisors, who will be invited to participate in the meeting and prompted to suggest further advisors, likely reaching 10-15 experts in total. The list of advisors will also be used for Stage 3 (real-time Delphi exercise). The expanded working group will make its best efforts to assure diversity, variety in career stages, geography, gender, race, and multicultural representation. The extended working group will also meet every other week, and each meeting will ideally be composed of 3-7 people, rotating participants, with at least one external advisor present (otherwise be rescheduled). The rotation and size limit of participants in a single meeting is built upon our prior experience with qualitative research, specifically focus groups, where compact size and diversified expertise aid to better reach data saturation.[32,33] The working group will work on: (a) review of existing EQUATOR/PCORI reporting guidelines related to prediction modeling and treatment effect estimation; (b) evaluation of published scoping reviews of counterfactual prediction modeling studies for biomedical sciences, and development of a new systematic review; (c) drafting of the initial reporting checklist for the Delphi survey; (d) review of the survey and development of the final guideline; (e) manuscript writing; and (f) submission of the products to peer-review, publication, and dissemination.

Stage 2: Literature review of counterfactual prediction modeling studies

The purpose of the literature review is twofold: (1) to build a knowledge base on study design, methodological approaches, use cases, and reporting commonalities among causal inference and counterfactual prediction studies in biomedical sciences; and (2) to help the development of reporting items for PRECOG. A subset of the working group members will concentrate on the review. In 2021, Lin et al.[34] published a scoping review on causal methods for predictions under hypothetical interventions, screening nearly 5,000 papers and focusing on 13 key articles, including traditional statistical as well as machine learning modeling. Most works used marginal structural models and g-computation. The authors concluded that "techniques for validating causal prediction models' are still in their infancy." Based on the results from the scoping review, and expanding the search strategy and the article sources, the team is going to move forward with a systematic review. The review will provide counts on methodology, review, and applied papers, but then will focus on works that include at least one observational data source and an application use case, further deepening the validation strategies. The planned reporting statement of choice is the "preferred reporting items for systematic reviews and meta-analyses" (PRISMA),[35] and the working group will register the work in the "prospective register of systematic reviews" (PROSPERO).[36]

As part of the review, we foresee discussing how to assess the potential risk of bias (which can lead to misuse and patients' harm), and if current tools such as "prediction model risk of bias assessment tool" (PROBAST) are appropriate.[37]

Stage 3: Real-time Delphi exercise

We will conduct a real-time Delphi survey[38] to review and refine the items of the PRECOG reporting checklist. Participants will be identified initially through the professional network of the core working group and of the external advisors, and further via literature search (including but not limited to the existing scoping review and the planned systematic review), social media screening, and snowballing by the active participants. As for the expanded working group composition, participants will be invited from diverse and multicultural backgrounds and different countries. Invitees will include academics at various career stages, researchers and investigators from non-profit and for-profit organizations, program officers from national/federal funding agencies, entrepreneurs, health care professionals, journal editors, policymakers, health care regulators, and end-users of predictive models. The participant selection will be based on area expertise grouping (computer science, biostatistics, biomedical informatics, statistics, epidemiology, standards, causal inference, ethics), used to determine the sample size (discussed below). We choose a computer-based, real-time Delphi,[38] since it offers some operational advantages with respect to conventional multi-round Delphi techniques, e.g., responder's attrition.[39] In brief, real-time Delphi is a "roundless" exercise based on an online survey platform. Participants can access and modify their responses at any time during the survey timeframe, and they can view the survey summaries calculated among all responders. In this way, participants can see if/how their opinion is unpopular and add further comments to support their cases.

The working group will develop an initial reporting checklist for PRECOG, based on the EQUATOR developing standard and existing related guidelines/statements. We anticipate that PRECOG will draw substantially from the reporting items of TRIPOD as well as the recommendations of PATH; however, we expect major differences rather than a simple merge. For instance, performance evaluation as recommended in TRIPOD should be modified to include specific metrics such as the Precision Estimation of Heterogeneous Effects (PEHE),[40] and emphasize out-of-distribution validation. Another important aspect is the causal assumptions. PATH relies on RCTs, where randomization supports the strong ignorability of treatment assignments, while PRECOG models might be exclusively built on observational data (or a mixture of observational and RCT data) and a justification for causal claims will need to be provided.

An anonymous online survey will be created where each checklist item can be evaluated in relation to its importance and relevance for the guideline, using a five-point Likert scale, and a free text box for comments. Also, at the end of the survey, another text box will allow more generic comments and propositions, e.g., new items to be added to the checklist. When a participant consents to participate and completes the survey for the first time, they can view the summary of all responses to date and can access the survey again within the next six weeks. The survey is closed after the required sample size is reached, or a maximum of six weeks are passed from the last recorded first response.

There is no consensus on the sample size of a Delphi panel but a minimum number of 10-18 panel members per area of expertise has been recommended.[41] We will aim to reach a minimum sample size of 60 considering the aforementioned background expertise areas, compiling a list of 80-100 potential participants for the recruitment. At the end of the Delphi survey, the expanded working group will review the results and consolidate the checklist through a consensus meeting. The workgroup will also decide on the consensus rule. In general, for items ranked on a five-point Likert scale, the consensus rule is 80%,[42] but there can be differences in how adjacent items are grouped or weighted toward consensus.[43] For instance, Naughton et al.[44] quantified the Likert points from 1 (most important) to 5 (least important), and defined consensus for items scoring a median of 2.5 or less overall, when at least 80% of responders gave 1 to 3 points. More recent works proposed entropy-based consensus.[45]

Stage 4: Development of the guideline and related products

Upon finalization of the reporting checklist from the Delphi exercise, the extended working group will develop the full PRECOG guidelines. The manuscript will be posted to a public pre-print website, e.g., bioRxiv or medRxiv, before submission to a peer-review journal, and possibly presented as an abstract/poster in major international conferences, e.g., the annual conference of the American Medical Informatics Association (AMIA) or the Society for Epidemiology Research (SER). It is expected that the PRECOG initiative will produce at least the following papers:

- Guideline development protocol (this work).
- A systematic review of causal and counterfactual prediction models in biomedical sciences.
- PRECOG guideline.

Stage 5: Publication and dissemination plan

After being posted on preprint servers, the aforementioned manuscripts will be submitted to peer-reviewed international journals for final publication. The authors' list will be determined based on effective individual contributions, following the "contributor roles taxonomy" (CRediT) (https://casrai.org/credit/), and might include additional contributors other than the working group members and external advisors. The dissemination strategy will be discussed during the workgroup meetings. In addition to conferences and publications, it is likely that social media platforms such as Twitter will be leveraged to inform on the PRECOG availability and utility.

Patient and public involvement

This study does not include patients. However, the participants of the working groups –by definition– will be involved in the design of the Delphi survey, in its evaluation, and in the finalization of the PRECOG guideline (including authorship in papers). The participants of the Delphi survey can provide not only an evaluation of items but suggest new ones and re-evaluate the items during the time when the survey is open.

ETHICS AND DISSEMINATION

The study will follow the principles of the Declaration of Helsinki. The study has been registered in EQUATOR (https://tinyurl.com/2p88ucnb) and approved by University of Florida's IRB (protocol no. IRB202200495). Informed consent will be obtained from both the working groups and the Delphi survey participants. The dissemination of PRECOG and its products will be done through journal publications, conferences, websites, and social media, based on discussions by the workgroup, as described above.

Acknowledgments We thank the TRIPOD co-authors Dr. G. Collins (U Oxford, UK) and Dr. K.G. Moons (UMC Utrecht, NL), and Dr. N. Peek (U Manchester, UK) for expressing their interest to join the PRECOG working groups.

Contributors JX wrote and submitted the protocol description. YG performed an initial literature review on reporting standards. FW and HX performed an initial literature review on counterfactual prediction models. RL advised on protocol procedures and ethical review. JB and MP conceived the idea.

Funding This work has been in part supported by National Institutes of Health (NIH) - National Institute of Allergy and Infectious Diseases (NIAID) grants no. R01AI145552 and R01AI141810 (Dr. Prosperi), by National Institute on Aging (NIA) grants no. R33AG062884-03 (Dr. Lucero and Dr. Prosperi) and 5R21AG068717-02 (Dr. Bian and Dr. Guo), by National Cancer Institute (NCI) grants no. 5R01CA246418-02, 3R01CA246418-02S1, 1R21CA245858-01A1, 3R21CA245858-01A1S1,and 1R21CA253394-01A1 (Dr. Bian and Dr. Guo), by the Centers for Disease Control and Prevention (CDC) grant no. U18DP006512 (Dr. Bian, Dr. Guo, and Dr. Prosperi), and by University of Florida Informatics Institute Seed grant.

Competing interests None declared.

References

1. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement. Vol. 67, European Urology. 2015. p. 1142–51.

2. Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. J Clin Epidemiol. 2021 Oct;138:219–26.

3. Collins GS, van Smeden M, Riley RD. COVID-19 prediction models should adhere to methodological and reporting standards. Eur Respir J. 2020 Sep;56(3).

4. Hernán MA, Hsu J, Healy B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. Vol. 32, CHANCE. 2019. p. 42–9.

5. Wilkinson J, Arnold KF, Murray EJ, van Smeden M, Carr K, Sippy R, et al. Time to reality check the promises of machine learning-powered precision medicine. Lancet Digit Health. 2020 Dec;2(12):e677–80.

6. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset Shift in Machine Learning. MIT Press;

2008. 248 p.

7. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Vol. 2, Nature Machine Intelligence. 2020. p. 369–75.

8. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Vol. 66, Journal of Educational Psychology. 1974. p. 688–701.

9. Pearl J, Glymour M, Jewell NP. Causal Inference in Statistics: A Primer. John Wiley & Sons; 2016. 160 p.

10. Dorresteijn JAN, Visseren FLJ, Ridker PM, Wassink AMJ, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. Vol. 343, BMJ. 2011. p. d5888–d5888.

11. Nguyen TL, Collins GS, Landais P, Le Manach Y. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials—An illustration with the International Stroke Trial. Journal of clinical epidemiology. 2020 Sep 1;125:47-56.

12. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci U S A. 2019 Mar 5;116(10):4156–65.

13. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. Vol. 30, Statistics in Medicine. 2011. p. 2867–80.

14. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics. 2012 Dec;68(4):1010–8.

15. Lamont A, Lyons MD, Jaki T, Stuart E, Feaster DJ, Tharmaratnam K, et al. Identification of predicted individual treatment effects in randomized clinical trials. Vol. 27, Statistical Methods in Medical Research. 2018. p. 142–57.

16. Brown K, Merrigan P, Royer J. Estimating Average Treatment Effects With Propensity Scores Estimated With Four Machine Learning Procedures: Simulation Results in High Dimensional Settings and With Time to Event Outcomes. SSRN Electronic Journal.

17. Hu L, Lin J-Y (Joyce), Sigel K, Kale M. Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis. Ann Epidemiol. 2021 Oct;62:36–42.

18. Xiong M. Deep Learning for Causal Inference. Artificial Intelligence and Causal Inference. 2022. p. 151–208.

19. Ghosh S, Boucher C, Bian J, Prosperi M. Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN). Comput Methods Programs Biomed Update. 2021 Jul 16;1.

20. Ge Q, Huang X, Fang S, Guo S, Liu Y, Lin W, et al. Conditional Generative Adversarial Networks for Individualized Treatment Effect Estimation and Treatment Selection. Front Genet. 2020 Dec 11;11:585804.

21. Lu M, Sadiq S, Feaster DJ, Ishwaran H. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. J Comput Graph Stat. 2018 Feb 1;27(1):209–19.

22. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Int J Surg. 2011 Oct 13;9(8):672–7.

23. Lee H, Cashin AG, Lamb SE, Hopewell S, Vansteelandt S, VanderWeele TJ, et al. A Guideline for Reporting Mediation Analyses of Randomized Trials and Observational Studies: The AGReMA Statement. JAMA. 2021 Sep 21;326(11):1045–56.

24. Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, et al. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization: The STROBE-MR Statement. JAMA. 2021 Oct 26;326(16):1614–21.

25. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ. 2018 Nov 14;363:k3532.

26. Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. Pharmacoepidemiol Drug Saf. 2010 Jun;19(6):537–54.

27. Seeger JD. Standards for Causal Inference Methods in Analyses of Data from Observational and Experimental Studies in Patient-Centered Outcomes Research. 2012.

28. Murray EJ, Swanson SA, Hernán MA. Guidelines for estimating causal effects in pragmatic randomized trials. arXiv preprint arXiv:1911.06030. 2019 Nov 14.

29. Baker SG. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020 Jun 2;172(11):775–6.

30. Moher D, Schulz KF, Simera I, Altman DG. Guidance for Developers of Health Research Reporting Guidelines. Vol. 7, PLoS Medicine. 2010. p. e1000217.

31. 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 Jul 9;11(7):e048008.

32. Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: A systematic review of empirical tests. Soc Sci Med. 2022 Jan;292:114523.

33. Rich SN, Richards VL, Mavian CN, Switzer WM, Rife Magalis B, Poschman K, et al. Employing Molecular Phylodynamic Methods to Identify and Forecast HIV Transmission Clusters in Public Health Settings: A Qualitative Study. Viruses. 2020 Aug 22;12(9).

34. Lin L, Sperrin M, Jenkins DA, Martin GP, Peek N. A scoping review of causal methods enabling predictions under hypothetical interventions. Diagn Progn Res. 2021 Feb 4;5(1):3.

35. 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. Rev Esp Cardiol. 2021 Sep;74(9):790–9.

36. Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Vol. 1, Systematic Reviews. 2012. Available from: http://dx.doi.org/10.1186/2046-4053-1-2

37. 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 Jan 1;170(1):W1–33.

38. Gordon T, Pease A. RT Delphi: An efficient, “round-less” almost real time Delphi method. Vol. 73, Technological Forecasting and Social Change. 2006. p. 321–33.

39. Hall DA, Smith H, Heffernan E, Fackrell K, for the Core Outcome Measures. Recruiting and retaining participants in e-Delphi surveys for core outcome set development: Evaluating the COMiT’ID study. PLoS One. 2018 Jul 30;13(7):e0201378.

40. Hill JL. Bayesian Nonparametric Modeling for Causal Inference. Vol. 20, Journal of Computational and Graphical Statistics. 2011. p. 217–40.

41. Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. Vol. 42, Information & Management. 2004. p. 15–29.

42. Lynn MR. Determination and Quantification Of Content Validity. Vol. 35, Nursing Research. 1986.

43. Hsu C-C, Sandford BA. The Delphi Technique: Making Sense of Consensus. Practical Assessment, Research, and Evaluation. 2007;12(1):10.

44. Naughton B, Roberts L, Dopson S, Brindley D, Chapman S. Medicine authentication technology as a counterfeit medicine-detection tool: a Delphi method study to establish expert opinion on manual medicine authentication technology in secondary care. Vol. 7, BMJ Open. 2017. p. e013838.

45. Tastle WJ, Wierman MJ. Consensus and dissention: A measure of ordinal dispersion. Vol. 45, International Journal of Approximate Reasoning. 2007. p. 531–45.

FIGURE TITLES

Figure 1. Flowchart of the development of the reporting guideline for causal and counterfactual prediction models (PRECOG)

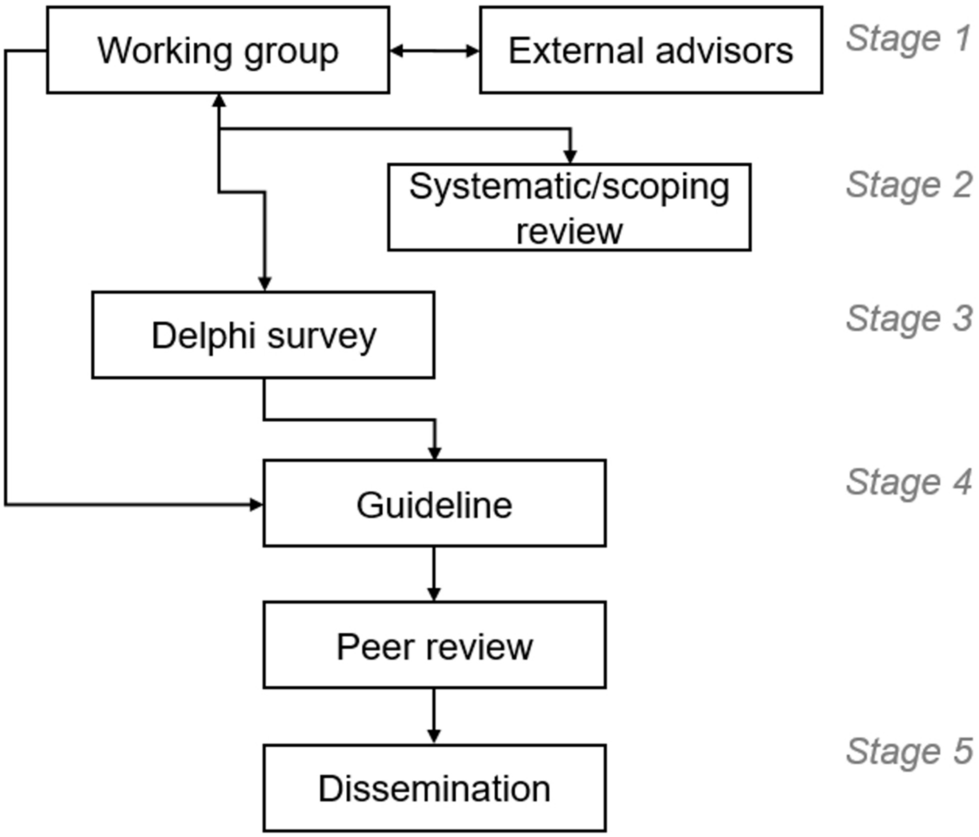

Figure 1. Flowchart of the development of the reporting guideline for causal and counterfactual prediction models (PRECOG).

109x93mm (765 x 765 DPI)