BMJ Open

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

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

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, Florida, USA
[2]Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York City, New York, USA
[3]School of Biomedical Informatics, University of Texas Health Science at Houston, Houston, Texas, USA
[4]School of Nursing, University of California - Los Angeles, Los Angeles, California, USA
[5]Department of Epidemiology, University of Florida, Gainesville, Florida, USA

Correspondence to
Mattia Prosperi;
m.prosperi@ufl.edu

ABSTRACT

Introduction While there are guidelines for reporting on observational studies (eg, Strengthening the Reporting of Observational Studies in Epidemiology, Reporting of Studies Conducted Using Observational Routinely Collected Health Data Statement), estimation of causal effects from both observational data and randomised experiments (eg, A Guideline for Reporting Mediation Analyses of Randomised Trials and Observational Studies, Consolidated Standards of Reporting Trials, PATH) and on prediction modelling (eg, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis), none is purposely made for deriving and validating models from observational data to predict counterfactuals for individuals on 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 (PRECOG).

Methods and analysis PRECOG will be developed following published guidance from the Enhancing the Quality and Transparency of Health Research (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 modelling for biomedical sciences (registered in Prospective Register of Systematic Reviews). 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 individualised effects of interventions.
⇒ Prediction of Counterfactuals Guideline (PRECOG) will fill a gap in reporting standards for counterfactual prediction modelling and will capitalise on the systematisation and quality of the Enhancing the Quality and Transparency of Health Research network.
⇒ PRECOG will be built on 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 on 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 optimisation purposes. Despite the availability of reporting guidelines, for example, '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, that is, 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, out-of-distribution populations (even when cross-validation performance is high)[6] and when trying to optimise outcomes.[7]

Thus, formal causal assessment is needed when developing prediction models on observational data to be used for alternative scenarios and interventions, that is, counterfactual prediction models. The approaches from traditional statistics, computational science and econometrics, including the potential outcomes framework,[8] do-calculus and directed acyclic graphs,[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 individualised 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 modelling studies, there is a need for common grounds on model reporting, to improve overall quality (although 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 randomised clinical trials (RCTs), for example, 'Consolidated Standards of Reporting Trials'[22] and 'A Guideline for Reporting Mediation Analyses of Randomised Trials and Observational Studies'.[23] Reporting guidelines for observational studies also mention causal effects inference, for example, 'Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomisation',[24] 'Reporting of Studies Conducted Using Observational Routinely Collected Health Data Statement for Pharmacoepidemiology'[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 randomised trials.[28] Worth noting is the 'predictive approaches to treatment effect heterogeneity' (PATH) statement,[29] which—although 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 (eg, machine learning models with many df) that could bring overfitting. Overall, existing guidelines are not well fitted for causal and counterfactual prediction modelling 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 individualised treatment/intervention effects in the form of potential outcomes. On the one hand, the development of these models can follow both risk-and-effect-modelling 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 optimisation (eg, 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 policy-makers to carry out and critically appraise causal and counterfactual prediction modelling studies. We anticipate further expansion of the guideline for specific areas, for example, 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 modelling 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-Artificial Intelligence (AI) and Prediction model Risk Of Bias ASsessment Tool (PROBAST)-AI.[31] The expected timeline for stages 1–4 is 1 year, using 6–9 months for stages 1–2, and 3–6 months for stages 3–4.

Stage 1: working group setup and meetings
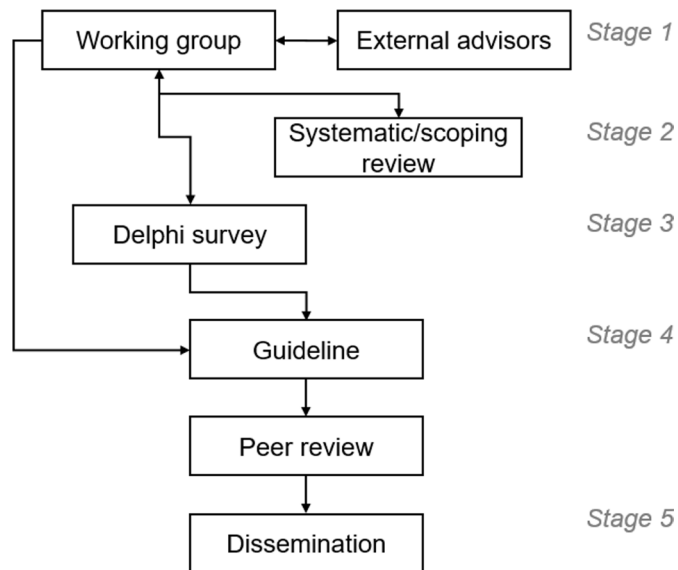The core working group is composed of the coauthors of this protocol description, who met every other week

Figure 1 Flow chart of the development of the reporting guideline for causal and counterfactual prediction models.

(30–45 min) since 13 September 2021, to discuss the development of the protocol itself, prepare documentation for the institutional review board, 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 on 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: (1) review of existing EQUATOR/PCORI reporting guidelines related to prediction modelling and treatment effect estimation; (2) evaluation of published scoping reviews of counterfactual prediction modelling studies for biomedical sciences and development of a new systematic review; (3) drafting of the initial reporting checklist for the Delphi survey; (4) review of the survey and development of the final guideline; (5) manuscript writing and (6) submission of the products to peer-review, publication and dissemination.

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 the development of reporting items for PRECOG. A subset of the working group members will concentrate on the review. Lin et al[34] published a scoping review on causal methods for predictions under hypothetical interventions, screening nearly 5000 papers and focusing on 13 key articles, including traditional statistical as well as machine learning modelling. 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',[35] and the working group will register the work in the 'Prospective Register of Systematic Reviews.[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 '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 organisations, programme officers from national/federal funding agencies, entrepreneurs, healthcare professionals, journal editors, policy-makers, healthcare 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, for example, 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 time frame, 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,[40] and emphasise out-of-distribution validation. Another important aspect is the causal assumptions. PATH relies on RCTs, where randomisation 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, for example, 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 6 weeks. The survey is closed after the required sample size is reached, or a maximum of 6 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 1–3 points. More recent works proposed entropy-based consensus.[45]

Stage 4: development of the guideline and related products
On finalisation 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 preprint website, for example, bioRxiv or medRxiv, before submission to a peer-review journal and possibly presented as an abstract/poster in major international conferences, for example, the annual conference of the American Medical Informatics Association or the Society for Epidemiology Research. 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 finalisation 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.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Jie Xu http://orcid.org/0000-0001-5291-5198
Yi Guo http://orcid.org/0000-0003-0587-4105
Robert Lucero http://orcid.org/0000-0002-8089-466X
Jiang Bian http://orcid.org/0000-0002-2238-5429

REFERENCES

1 Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol* 2015;67:1142–51.
2 Van Calster B, Wynants L, Riley RD, *et al*. Methodology over metrics: current scientific standards are a disservice to patients and society. *J Clin Epidemiol* 2021;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;56:2002643.
4 Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. *Chance* 2019;32:42–9.
5 Wilkinson J, Arnold KF, Murray EJ, *et al*. Time to reality check the promises of machine learning-powered precision medicine. *Lancet Digit Health* 2020;2:e677–80.
6 Quinonero-Candela J, Sugiyama M, Schwaighofer A. *Dataset shift in machine learning*. MIT Press, 2008: 248 p.
7 Prosperi M, Guo Y, Sperrin M, *et al*. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat Mach Intell* 2020;2:369–75.
8 Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 1974;66:688–701.
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, *et al*. Estimating treatment effects for individual patients based on the results of randomised clinical trials. Vol. 343. *BMJ* 2011:d5888–d5888.
11 Nguyen T-L, Collins GS, Landais P, *et al*. Counterfactual clinical prediction models could help to infer individualized treatment effects in randomized controlled trials-An illustration with the International stroke trial. *J Clin Epidemiol* 2020;125:47–56.
12 Künzel SR, Sekhon JS, Bickel PJ, *et al*. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci U S A* 2019;116:4156–65.
13 Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Stat Med* 2011;30:2867–80.
14 Zhang B, Tsiatis AA, Laber EB, *et al*. A robust method for estimating optimal treatment regimes. *Biometrics* 2012;68:1010–8.
15 Lamont A, Lyons MD, Jaki T, *et al*. Identification of predicted individual treatment effects in randomized clinical trials. *Stat Methods Med Res* 2018;27:142–57.
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*;72.
17 Hu L, Lin J-Y, Sigel K, *et al*. Estimating heterogeneous survival treatment effects of lung cancer screening approaches: a causal machine learning analysis. *Ann Epidemiol* 2021;62:36–42.
18 Xiong M. Deep learning for causal inference. *Artificial Intelligence and Causal Inference* 2022:151–208.
19 Ghosh S, Boucher C, Bian J, *et al*. Propensity score synthetic augmentation matching using generative adversarial networks (PSSAM-GAN). *Comput Methods Programs Biomed Update* 2021;1. doi:10.1016/j.cmpbup.2021.100020. [Epub ahead of print: 16 07 2021].
20 Ge Q, Huang X, Fang S, *et al*. Conditional generative Adversarial networks for individualized treatment effect estimation and treatment selection. *Front Genet* 2020;11:585804.
21 Lu M, Sadiq S, Feaster DJ, *et al*. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat* 2018;27:209–19.
22 Schulz KF, Altman DG, Moher D, *et al*. Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. *Int J Surg* 2011;9:672–7.
23 Lee H, Cashin AG, Lamb SE, *et al*. A guideline for reporting mediation analyses of randomized trials and observational studies: the AGReMA statement. *JAMA* 2021;326:1045–56.
24 Skrivankova VW, Richmond RC, Woolf BAR, *et al*. Strengthening the reporting of observational studies in epidemiology using Mendelian randomization: the STROBE-MR statement. *JAMA* 2021;326:1614–21.
25 Langan SM, Schmidt SA, Wing K, *et al*. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* 2018;363:k3532.
26 Brookhart MA, Rassen JA, Schneeweiss S. Instrumental variable methods in comparative safety and effectiveness research. *Pharmacoepidemiol Drug Saf* 2010;19:537–54.
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.
29 Baker SG. The predictive approaches to treatment effect heterogeneity (path) statement. *Ann Intern Med* 2020;172:775–6.
30 Moher D, Schulz KF, Simera I, *et al*. Guidance for developers of health research reporting guidelines. *PLoS Med* 2010;7:p. e1000217.
31 Collins GS, Dhiman P, Andaur Navarro CL, *et al*. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.
32 Hennink M, Kaiser BN. Sample sizes for saturation in qualitative research: a systematic review of empirical tests. *Soc Sci Med* 2022;292:114523.
33 Rich SN, Richards VL, Mavian CN, *et al*. Employing molecular Phylodynamic methods to identify and forecast HIV transmission clusters in public health settings: a qualitative study. *Viruses* 2020;12. doi:10.3390/v12090921. [Epub ahead of print: 22 08 2020].
34 Lin L, Sperrin M, Jenkins DA, *et al*. A scoping review of causal methods enabling predictions under hypothetical interventions. *Diagn Progn Res* 2021;5:3.
35 Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Rev Esp Cardiol* 2021;74:790–9.
36 Booth A, Clarke M, Dooley G, *et al*. The nuts and bolts of Prospero: an international prospective register of systematic reviews. Vol. 1, systematic reviews. 2012. Available: http://dx.doi.org/10.1186/2046-4053-1-2
37 Moons KGM, Wolff RF, Riley RD, *et al*. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019;170:W1–33.
38 Gordon T, Pease A. RT Delphi: An efficient, "round-less" almost real time Delphi method. *Technol Forecast Soc Change* 2006;73:321–33.
39 Hall DA, Smith H, Heffernan E, *et al*. Recruiting and retaining participants in e-Delphi surveys for core outcome set development: evaluating the COMiT'ID study. *PLoS One* 2018;13:e0201378.
40 Hill JL. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 2011;20:217–40.
41 Okoli C, Pawlowski SD. The Delphi method as a research tool: an example, design considerations and applications. *Information & Management* 2004;42:15–29.
42 Lynn MR. Determination and quantification of content validity. *Nurs Res* 1986;35:382–6.
43 Hsu C-C, Sandford BA. The Delphi technique: making sense of consensus. *Practical Assessment, Research, and Evaluation* 2007;12:10.
44 Naughton B, Roberts L, Dopson S, *et al*. 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. *BMJ Open* 2017;7:p. e013838.
45 Tastle WJ, Wierman MJ. Consensus and dissention: a measure of ordinal dispersion. *International Journal of Approximate Reasoning* 2007;45:531–45.