RESEARCH

Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses

Stefan Schandelmaier MD PhD, Matthias Briel MD PhD, Ravi Varadhan PhD, Christopher H. Schmid PhD, Niveditha Devasenapathy MBBS MSc, Rodney A. Hayward MD, Joel Gagnier MD PhD, Michael Borenstein PhD, Geert J.M.G. van der Heijden PhD, Issa J. Dahabreh MD PhD, Xin Sun PhD, Willi Sauerbrei PhD, Michael Walsh MD PhD, John P.A. Ioannidis MD DSc, Lehana Thabane PhD, Gordon H. Guyatt MD MSc

Cite as: *CMAJ* 2020 August 10;192:E901-6. doi: 10.1503/cmaj.200077

ABSTRACT

BACKGROUND: Most randomized controlled trials (RCTs) and meta-analyses of RCTs examine effect modification (also called a subgroup effect or interaction), in which the effect of an intervention varies by another variable (e.g., age or disease severity). Assessing the credibility of an apparent effect modification presents challenges; therefore, we developed the Instrument for assessing the Credibility of Effect Modification Analyses (ICEMAN).

METHODS: To develop ICEMAN, we established a detailed concept; identified candidate credibility considerations in a systematic survey of the literature; together with experts, performed a consensus study to identify key considerations and develop them into instrument items; and refined the instrument based on feedback from trial investigators, systematic review authors and journal editors, who applied drafts of ICEMAN to published claims of effect modification.

RESULTS: The final instrument consists of a set of preliminary considerations, core questions (5 for RCTs, 8 for meta-analyses) with 4 response options, 1 optional item for additional considerations and a rating of credibility on a visual analogue scale ranging from very low to high. An accompanying manual provides rationales, detailed instructions and examples from the literature. Seventeen potential users tested ICEMAN; their suggestions improved the user-friendliness of the instrument.

INTERPRETATION: The Instrument for assessing the Credibility of Effect Modification Analyses offers explicit guidance for investigators, systematic reviewers, journal editors and others considering making a claim of effect modification or interpreting a claim made by others.

Investigators who conduct randomized controlled trials (RCTs) and meta-analyses of RCTs often perform analyses of effect modification to assess whether intervention effects might vary by another variable such as age, disease severity or, in a meta-analysis, study setting or year of study.[1–14] The terminology varies; Box 1 presents the alternatives currently in use.

Investigators sometimes make claims that an effect modification is present. Literature surveys suggest that 14%–26% of RCTs and meta-analyses emphasize at least 1 potential effect modification in their abstract or discussion.[4–9,11]

The interest in effect modification is understandable: if patients with differing characteristics respond differently to the same intervention, the overall effect estimate is misleading for some, if not all, patients. Identifying situations in which true variation in effects exist is important, and the notion of tailoring therapy to patients has enormous appeal. Moreover, the opportunities for analyzing effect modification grow with the increasing number of newly developed diagnostic and genomic markers.

However, mistaken claims of effect modification may compromise optimal patient care, and many claims of effect modification have subsequently proved spurious.[15–24] Applying a mistaken claim of effect modification may cause harms through administration of ineffective treatment or may lead to patients' being denied beneficial therapies, and will likely increase health care costs.

Numerous theoretical analyses and simulation studies show that the fundamental reason for misleading claims of effect

Box 1: Variation in terminology of effect modification

Synonyms for effect modification

- Subgroup effect
- Statistical interaction
- Moderation
- Differential effect
- Heterogeneity of treatment effects

Synonyms for effect modifier

- Subgrouping variable
- Predictor of treatment response
- Moderator

modification is chance:[25] even if the treatment effect is the same for all patients, examining a sufficiently large number of candidates will inevitably reveal an apparent, but misleading, effect modification. Other reasons that contribute to spurious claims include selective reporting,[5,7] lack of background knowledge and prior evidence,[5,7,26] and failure to use a proper statistical analysis.[5,8,27–29]

Nevertheless, some claims of effect modification — likely a minority[1] — will be true. Because most claims will never undergo replication to determine their veracity, stakeholders, including health care providers, clinical investigators, systematic review authors, guideline developers and journal editors, need criteria to differentiate spurious from real claims.

Methodologists first suggested credibility criteria for effect modification in the early 1990s.[30,31] Since then, 30 groups have suggested sets of 3–21 criteria.[25] Aside from the variability in these criteria, previous sets have suboptimal presentation, which results in ambiguity in their application.[25] Some criteria — for instance, whether the effect modifier was one of a small number tested[32] — are subjective, and users could benefit from more detailed guidance. Most important, none of the previous sets of criteria involved a rigorous development process or underwent user testing before publication.[25]

To address these limitations, we developed the Instrument for assessing the Credibility of Effect Modification Analyses (ICEMAN).

Methods

Development of ICEMAN consisted of 4 steps:[33,34] clarifying the scope and measurement concept of the instrument; a systematic literature survey to identify existing instruments and candidate credibility criteria; a consensus study among experts to identify key criteria and design the instrument; and formal user testing.

Concept

Members of the core group (S.S., G.H.G., M.B., X.S., M.W., L.T.) began with the following concept:

- Effect modification means that the effect of an intervention on an outcome varies by levels of another variable.
- The aim of the new instrument is to assist users in assessing the credibility of claims that effect modification is present (rather than claims that effect modification is absent, which would require different criteria).
- An effect modification is credible if it is unlikely to be the result of chance or bias.
- We also clarified that patient importance is not part of the credibility; an effect modification primarily defines an association between the modifier and the causal effect of the intervention on the outcome (i.e., the presence of a causal relation between the modifier and the outcome is not necessary for the claim to be valid); and effect modification can be assessed on any scale (e.g., risk ratio or risk difference).
- Target users include health care providers, trial investigators, systematic review authors, health technology assessment practitioners, journal editors, guideline developers and health policy-makers.
- The instrument will address individual RCTs and meta-analyses of RCTs (including aggregate data and individual participant data meta-analyses).
- The core instrument will consist of no more than 8–12 questions to keep both the demands of application and the cognitive burden manageable and will provide explicit response options for each question.
- Responses to individual criteria should vary when applied to different claims of effect modification. An overly strict or lenient criterion that does not vary is useless for distinguishing more from less credible claims.
- The instrument should conclude with a summary assessment that expresses the overall credibility of the apparent effect modification.

Systematic literature survey

The objectives of the systematic literature survey, presented in a separate publication,[25] were to identify existing instruments for assessing the credibility of effect modification, candidate credibility criteria and leading experts in the field. Based on a comprehensive search of journal articles and textbooks, we identified 150 eligible publications, from which we abstracted 36 candidate criteria (Appendix 1, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.200077/-/DC1), 30 previous sets of criteria (none reflected our concept sufficiently) and 40 experts.[25]

Consensus study

The aim of the consensus study was to identify key criteria to assess the credibility of effect modification claims and use them to develop a user-friendly instrument. The steering committee randomized the order of the 40 experts identified in the literature survey and invited the first 18 to participate, of whom 9 agreed, 6 declined, and 3 did not respond. The final group included 15 members (the core group and 9 experts: R.V., C.H.S., R.A.H., J.G., M.B., G.J.M.G.V., I.J.D., W.S. and J.P.A.I.). The consensus study included elements of the Delphi method complemented by interactive video conferences.

In a first step, S.S. created a list of the 36 candidate criteria identified in the systematic survey, their frequency of reporting and reported rationales for their use (Appendix 2, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.200077/-/DC1). The

members of the group (excluding S.S.) independently rated the importance of each criterion for credibility assessment from 1 (not important at all) to 7 (highly important). They also provided written suggestions to drop criteria, merge criteria or add new criteria. During the first video conference, the group discussed the importance ratings and identified 20 criteria that should be included (some of which we later combined), 8 that should be excluded and 8 that were considered optional (Appendix 2).

Based on the initial criteria selected, the core group developed a first draft of the instrument. Initially, we planned to create a single instrument applicable to individual RCTs, aggregate data meta-analyses and meta-analyses of individual participant data. Because a single version proved excessively complex, the group decided to create separate versions for RCTs and meta-analyses (of any type). We drafted preliminary considerations, explicit items (each with 4 response options), optional considerations and a final item to assess overall credibility by means of a visual analogue scale. Where possible, we used a format similar to that of other research assessment instruments such as the Cochrane risk-of-bias tool[35] and Grading of Recommendations Assessment, Development and Evaluation (GRADE).[36]

We held a second video-conference to reach consensus on the general structure of the instrument, including preliminary considerations, core items and format of the overall rating. Main discussion points included issues of threshold selection (e.g., for *p* values and number of analyses) and framing of optional considerations.

In the last part of the consensus study, we created a detailed manual that provides, for each item, a detailed justification of the importance of the item for assessing the credibility of effect modification claims (Appendix 3, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.200077/-/DC1) (the justifications for excluding candidate items are provided in Appendix 2). For each response option, we sought a supporting example of an effect modification claim from the medical literature.

Throughout the consensus study, we periodically circulated summaries of the discussions and updated versions of the instrument to the experts, inviting them to provide comments. Appendix 2 documents major developments.

User testing

The aim of user testing was to identify challenges experienced by potential users in applying ICEMAN to a claim of effect modification that we provided. Each user received the full text of an RCT or meta-analysis in which the authors claimed an effect modification, and drafts of ICEMAN and the manual. We selected 17 claims specifically to introduce variation across the range of possible credibility (4 very low, 5 low, 4 moderate and 4 high, based on our judgement) and across designs (9 RCTs, 4 aggregate data meta-analyses and 4 meta-analyses of individual participant data) (Appendix 4, Supplemental Table S1, available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.200077/-/DC1).

We recruited 17 potential users from 3 sources: corresponding authors of Cochrane reviews (n = 7), authors of published RCTs (n = 5) and journal editors from personal networks (n = 5). The users varied with respect to gender, background and familiarity with issues of effect modification (Appendix 4, Supplemental Table S2). We continued to enrol users until they did not identify any new major limitations of the instrument.

One of 2 investigators (S.S. or N.D.) interviewed users immediately after they had applied ICEMAN. The investigators followed a semistructured interview guide that included open questions (e.g., "What was your experience when you applied the first item?") and allowed expansion on topics that emerged during the interview. The interviews lasted 25–70 (median 37) minutes and were video-recorded. The interviewers transcribed the interviews and extracted suggestions for improvement using Dedoose software (www.dedoose.com). We updated the instrument after 7, 12 and 15 interviews, before the consensus group finalized the instrument and manual. The users' comments and resulting changes are summarized in Appendix 4, Supplemental Table S3.

Ethics approval

The Hamilton Integrated Research Ethics Board approved the user-testing study.

Results

The version of ICEMAN for individual RCTs is presented in Appendix 5 and that for meta-analyses of RCTs in Appendix 6 (both available at www.cmaj.ca/lookup/suppl/doi:10.1503/cmaj.200077/-/DC1). The material, including potential updates, can also be downloaded from https://www.iceman.help.

The instrument can be used by investigators performing RCTs or meta-analyses who are planning analyses of effect modification; by investigators evaluating the credibility of claims they are considering; and by those who are critically appraising effect modification claims in the literature.

Both versions start with a set of preliminary considerations that link ICEMAN to a specific study, specify the effect modification claim under consideration and alert users that ICEMAN may not apply to effect modifiers measured after randomization (see manual [Appendix 3] for more details).

The version for RCTs includes 5 core questions and that for meta-analyses, 8 core questions (4 questions overlap) (Table 1). For each core question, ICEMAN provides 4 response options that differ in wording but have the same order and logic: response options on the left indicate definitely or probably reduced credibility, and response options on the right indicate probably or definitely increased credibility. We included the response option "probably no" with "unclear" to cover situations with insufficient information.

One optional question allows additional considerations that can reduce or increase credibility, such as results from sensitivity analyses, a dose–response relation, or other considerations that are difficult to ascertain, are less relevant or do not universally apply (see manual [Appendix 3] for examples).

The instrument concludes with an overall credibility assessment rated on a visual analogue scale divided into 4 areas (very low credibility, low credibility, moderate credibility and high credibility). The 4 areas correspond roughly to probabilities of less than 25%, 25%–50%, 51%–75% and greater than 75%, respectively, that the effect modification truly exists. To

Table 1: Comparison of the core questions of the 2 versions of the Instrument for assessing the Credibility of Effect Modification Analyses

Core question	Version; question no.*	
	Randomized controlled trials	Meta-analyses
Is the analysis of effect modification based on comparison within rather than between trials?	–	1
For within-trial comparisons, is the effect modification similar from trial to trial?	–	2
For between-trial comparisons, is the number of trials large?	–	3
Was the direction of effect modification correctly hypothesized a priori?	1	4
Was the effect modification supported by prior evidence?	2	–
Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification?	3	5
Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?	4	6
Did the authors use a random-effects model?	–	7
If the effect modifier is a continuous variable, were arbitrary cut points avoided?	5	8

NA = not applicable.
*Numbers reflect order of appearance within the full instrument (see Appendices 5 and 6).

aid interpretation, the final item provides suggestions — rather than an algorithm — for judging overall credibility.

The manual (Appendix 3) provides detailed explanations, key references, examples for each response option, suggestions for use and presentation, and elaboration on conceptual considerations.

Interpretation

Using a systematic approach involving both experts and potential users, we developed ICEMAN. The instrument provides versions for individual RCTs and meta-analyses, is short (5 core items for RCTs, 8 for meta-analyses), is structured (preliminary considerations, core questions, overall rating) and provides a detailed manual.

One of the benefits of ICEMAN is that it may help to reduce overreliance on the *p* value for interaction when assessing credibility. The *p* value counts no more than other items. Instead of a single threshold, the response options are based on 3 thresholds, thus emphasizing the continuous character of the concept. The expectation is therefore that claims of effect modification can be reasonably credible despite borderline *p* values, whereas claims that are based exclusively on very low *p* values may have low credibility.

Limitations

A possible limitation of ICEMAN is that, to optimize reliability, formulating 4 response options required specification of threshold values for credibility with respect to the number of studies in a meta-analysis, *p* values and number of candidate effect modifiers. These thresholds are arbitrary, and experts initially disagreed on the specific threshold values and whether they should be used at all. Particularly controversial within our group were thresholds for interaction *p* values, although the group finally found a compromise acceptable to all (using thresholds of 0.05, 0.01 and 0.005). It is perhaps reassuring that none of the participants of the user-testing study mentioned concerns with the chosen thresholds, and those who did comment appreciated their explicitness. Nevertheless, some users may disagree with the chosen thresholds.

Another potential limitation is that the core questions may not include all credibility considerations that experienced analysts may deem relevant, in particular for complex analyses such as modelling of continuous effect modifiers[37,38] and data-driven algorithms for subgroup discovery.[39,40] For instance, some analysts may question the appropriateness of statistical models underlying tests for interaction,[26,41,42] may differ in their approach to multiple testing,[43] may consider 3-way or 4-way interactions, may correct for exaggerated magnitude of effect modification[44] or may want to consider the correlation structure between multiple effect modifiers.[45] Even for such users, ICEMAN will provide a useful starting point. For instance, if the core questions suggest low or very low credibility, it is unlikely that investing in additional, more complex analyses could increase credibility substantially; however, if the core questions suggest moderate credibility, users can use ICEMAN's optional item to incorporate additional considerations.

Some properties of ICEMAN remain uncertain. We plan to investigate the reliability of ICEMAN ratings when applied by different raters to claims of effect modification. Another open question is the validity of ICEMAN ratings. We are unsure, however, whether there will ever be sufficient data available to investigate validity if we consider independent replication the reference standard for establishing the credibility of an effect modification claim. A recent analysis showed that attempts to

replicate effect modification findings in RCTs are extremely rare.[24] Therefore, we invite ICEMAN users to share their ratings with us so we can start building a database of more or less credible claims of effect modification and, at a later time, potentially assess the extent to which the claims were replicated. This will also allow better calibration of the 4 credibility areas of the overall credibility assessment and the corresponding ranges of percent credibility that we suggest. In addition, we will continue to evaluate ICEMAN's performance in practice. We invite users to report difficulties or suggestions for improvement for consideration in future modifications of the instrument, by contacting the corresponding author.

Conclusion

In summary, ICEMAN provides a systematically developed and thoroughly user-tested instrument for judging the credibility of apparent effect modification. We expect that both investigators and readers of RCTs and meta-analyses, and other groups including journal editors, will find the structured assessment of credibility of proposed effect modification helpful.

References

1. Schuit E, Li AH, Ioannidis JPA. How often can meta-analyses of individual-level data individualize treatment? A meta-epidemiologic study. *Int J Epidemiol* 2019;48:596-608.
2. Gabler NB, Duan N, Raneses E, et al. No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals. *Trials* 2016;17:320.
3. Simmonds M, Stewart G, Stewart L. A decade of individual participant data meta-analyses: a review of current practice. *Contemp Clin Trials* 2015;45:76-83.
4. Zhang S, Liang F, Li W, et al. Subgroup analyses in reporting of phase III clinical trials in solid tumors. *J Clin Oncol* 2015;33:1697-702.
5. Donegan S, Williams L, Dias S, et al. Exploring treatment by covariate interactions using subgroup analysis and meta-regression in Cochrane reviews: a review of recent practice. *PLoS One* 2015;10:e0128804.
6. Barton SP, Peckitt C, Sclafani F, et al. The influence of industry sponsorship on the reporting of subgroup analyses within phase III randomised controlled trials in gastrointestinal oncology. *Eur J Cancer* 2015;51:2732-9.
7. Kasenda B, Schandelmaier S, Sun X, et al.; Drosoph Inf ServCO Study Group. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications [published erratum in *BMJ* 2014;349:4921]. *BMJ* 2014;349:g4539.
8. Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. *BMJ* 2012;344:e1553.
9. Wang R, Lagakos SW, Ware JH, et al. Statistics in medicine — reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007;357:2189-94.
10. Koopman L, van der Heijden GJ, Glasziou PP, et al. A systematic review of analytical methods used to study subgroups in (individual patient data) meta-analyses. *J Clin Epidemiol* 2007;60:1002-9.
11. Hernández AV, Boersma E, Murray GD, et al. Subgroup analyses in therapeutic cardiovascular clinical trials: Are most of them misleading? *Am Heart J* 2006; 151:257-64.
12. Bhandari M, Devereaux PJ, Li P, et al. Misuse of baseline comparison tests and subgroup analyses in surgical trials. *Clin Orthop Relat Res* 2006;447:247-51.
13. Moreira ED Jr, Stein Z, Susser E. Reporting on methods of subgroup analysis in clinical trials: a survey of four scientific journals. *Braz J Med Biol Res* 2001;34: 1441-6.
14. Assmann SF, Pocock SJ, Enos LE, et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet* 2000;355:1064-9.
15. Fields WS, Lemak NA, Frankowski RF, et al. Controlled trial of aspirin in cerebral ischemia. *Stroke* 1977;8:301-14.
16. Canadian Cooperative Study Group. A randomized trial of aspirin and sulfinpyrazone in threatened stroke. *N Engl J Med* 1978;299:53-9.
17. Collaborative overview of randomised trials of antiplatelet therapy — I: Prevention of death, myocardial infarction, and stroke by prolonged antiplatelet therapy in various categories of patients. Antiplatelet Trialists' Collaboration [published erratum in *BMJ* 1994;308:1540]. *BMJ* 1994;308:81-106.
18. Amery A, Birkenhäger W, Brixko P, et al. Influence of antihypertensive drug treatment on morbidity and mortality in patients over the age of 60 years. European Working Party on High blood pressure in the Elderly (EWPHE) results: subgroup analysis on entry stratification. *J Hypertens Suppl* 1986;4:S642-7.
19. Gueyffier F, Bulpitt C, Boissel JP, et al. Antihypertensive drugs in very old people: a subgroup meta-analysis of randomised controlled trials. INDANA Group. *Lancet* 1999;353:793-6.
20. Weisberg LA. The efficacy and safety of ticlopidine and aspirin in non-whites: analysis of a patient subgroup from the Ticlopidine Aspirin Stroke Study. *Neurology* 1993;43:27-31.
21. Gorelick PB, Richardson D, Kelly M, et al.; African American Antiplatelet Stroke Prevention Study Investigators. Aspirin and ticlopidine for prevention of recurrent stroke in black patients: a randomized trial. *JAMA* 2003;289:2947-57.
22. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005;365: 176-86.
23. Guyatt G. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. 3rd ed. New York: McGraw-Hill Education; 2015.
24. Wallach JD, Sullivan PG, Trepanowski JF, et al. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. *JAMA Intern Med* 2017;177:554-60.
25. Schandelmaier S, Chang Y, Devasenapathy N, et al. A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or meta-analyses. *J Clin Epidemiol* 2019;113:159-67.
26. Dahabreh IJ, Hayward R, Kent DM. Using group data to treat individuals: understanding heterogeneous treatment effects in the age of precision medicine and patient-centred evidence. *Int J Epidemiol* 2016;45:2184-93.
27. Koopman L, van der Heijden GJ, Hoes AW, et al. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. *Int J Technol Assess Health Care* 2008;24:358-61.
28. Fisher DJ, Carpenter JR, Morris TP, et al. Meta-analytical methods to identify who benefits most from treatments: Daft, deluded, or deft approach? *BMJ* 2017;356:j573.
29. Berlin JA, Santanna J, Schmid CH, et al.; Anti-Lymphocyte Antibody Induction Therapy Study Group. Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head. *Stat Med* 2002;21:371-87.
30. Yusuf S, Wittes J, Probstfield J, et al. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991;266: 93-8.
31. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med* 1992;116:78-84.
32. Sun X, Briel M, Walter SD, et al. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ* 2010;340:c117.
33. Streiner DL, Norman GR, Cairney J. *Health measurement scales: a practical guide to their development and use*. Oxford (UK): Oxford University Press; 2015.
34. Whiting P, Wolff R, Mallett S, et al. A proposed framework for developing quality assessment tools. *Syst Rev* 2017;6:204.
35. Higgins J, Sterne J, Savović J, et al. A revised tool for assessing risk of bias in randomized trials. *Cochrane Database Syst Rev* 2016;(Suppl 1):29-31.
36. Alonso-Coello P, Schünemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ* 2016;353:i2016.
37. Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. *Stat Med* 2013;32:3788-803.
38. Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis. *Stat Med* 2014;33: 4695-708.
39. Lipkovich I, Dmitrienko A, D'Agostino BR Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med* 2017;36: 136-96.
40. Lipkovich I, Dmitrienko A, Muysers C, et al. Multiplicity issues in exploratory subgroup analysis. *J Biopharm Stat* 2018;28:63-81.
41. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Stat Med* 2004;23:2509-25.
42. Borenstein M, Higgins JP. Meta-analysis and subgroups. *Prev Sci* 2013;14: 134-43.
43. Dmitrienko A, D'Agostino R Sr. Traditional multiplicity adjustment methods in clinical trials. *Stat Med* 2013;32:5172-218.
44. Varadhan R, Wang SJ. Treatment effect heterogeneity for univariate subgroups in clinical trials: shrinkage, standardization, or else. *Biom J* 2016;58:133-53.
45. Varadhan R, Wang SJ. Standardization for subgroup analysis in randomized controlled trials. *J Biopharm Stat* 2014;24:154-67.

Competing interests: None declared.

This article has been peer reviewed.

Affiliations: Departments of Health Research Methods, Evidence, and Impact (Schandelmaier, Briel, Walsh, Thabane, Guyatt), Medicine (Walsh, Guyatt), Pediatrics (Thabane) and Anesthesia (Thabane), McMaster University, Hamilton, Ont.; Institute for Clinical Epidemiology and Biostatistics (Schandelmaier, Briel), Department of Clinical Research, Basel University, Basel, Switzerland; Division of Biostatistics and Bioinformatics (Varadhan), Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, Md.; Department of Biostatistics (Schmid), Brown University School of Public Health, Brown University, Providence, RI; Indian institute of Public Health-Delhi (Devasenapathy), Public Health Foundation of India, New Delhi, India; VA Center for Clinical Management and Research (Hayward); Department of Internal Medicine (Hayward), University of Michigan School of Medicine; Department of Orthopaedic Surgery (Gagnier), University of Michigan; Department of Epidemiology (Gagnier), School of Public Health, University of Michigan, Ann Arbor, Mich.; Biostat (Borenstein), Englewood, NJ; Department of Social Dentistry (van der Heijden), Academic Center for Dentistry Amsterdam, University of Amsterdam and VU University Amsterdam, Amsterdam, Netherlands; Center for Evidence Synthesis in Health (Dahabreh) and Departments of Health Services, Policy, and Practice (Dahabreh) and Epidemiology (Dahabreh), School of Public Health, Brown University, Providence, RI; Chinese Evidence-Based Medicine Center (Sun), West China Hospital, Sichuan University, Chengdu, China; Institute of Medical Biometry and Statistics (Sauerbrei), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany; Population Health Research Institute (Walsh), Hamilton Health Sciences/McMaster University, Hamilton, Ont.; Departments of Medicine (Ioannidis), Health Research and Policy (Ioannidis) and Biomedical Data Science (Ioannidis), and Statistics and Meta-Research Innovation Center at Stanford (METRICS) (Ioannidis), Stanford University, Stanford, Calif.; Biostatistics Unit (Thabane), St. Joseph's Healthcare, Hamilton, Ont.

Contributors: Stefan Schandelmaier, Matthias Briel, Xin Sun, Michael Walsh, Lehana Thabane and Gordon Guyatt conceived the study. Stefan Schandelmaier and Gordon Guyatt drafted the manuscript. All of the authors revised the manuscript critically for important intellectual content, gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Funding: Stefan Schandelmaier was supported by grant P300PB_16475 from the Swiss National Science Foundation, the Gottfried and Julia Bangerter-Rhyner Foundation and the Freiwillige Akademische Gesellschaft Basel. Issa Dahabreh was supported by Methods Research Awards ME-1306-03758 and ME-1502-27794 from the Patient-Centered Outcomes Research Institute.

Disclaimer: The content of this manuscript does not represent the official views of the Patient-Centered Outcomes Research Institute, its Board of Governors or the Methodology Committee.

Data sharing: All relevant data except the names of the participants in the user-testing study and the transcripts of the interviews are provided in the publication, appendices and a related publication, and are available for use by other researchers. Additional requests regarding data sharing may be directed to the corresponding author.

Accepted: Apr. 6, 2020

Correspondence to: Stefan Schandelmaier, s.schandelmaier@gmail.com