Journal of Clinical Epidemiology 113 (2019) 159–167

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or meta-analyses

Stefan Schandelmaier[a,b,*], Yaping Chang[a], Niveditha Devasenapathy[c], Tahira Devji[a], Joey S.W. Kwong[d], Luis E. Colunga Lozano[a], Yung Lee[a,e], Arnav Agarwal[f], Neera Bhatnagar[a], Hannah Ewald[b], Ying Zhang[a,g], Xin Sun[h], Lehana Thabane[a,i], Michael Walsh[a,j], Matthias Briel[a,b], Gordon H. Guyatt[a,j]

[a]*Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[b]*Department of Clinical Research, Basel Institute for Clinical Epidemiology and Biostatistics, University of Basel and University Hospital Basel, Spitalstrasse 12, 4056 Basel, Switzerland*
[c]*Indian Institute of Public Health-Delhi, Public Health Foundation of India, Plot 47, Sector 44, Institutional Area, Gurgaon, 122002 Haryana, India*
[d]*JC School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong*
[e]*Michael G. DeGroote School of Medicine, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada*
[f]*Department of Medicine, University of Toronto, 190 Elizabeth Street, R. Fraser Elliott Building, 3-805, Toronto, Ontario M5G 2C4, Canada*
[g]*Center for Evidence-based Chinese Medicine, Beijing University of Chinese Medicine, 11 Bei San Huan Dong Lu, Chaoyang, Beijing 100029, China*
[h]*Chinese Evidence-Based Medicine Center, West China Hospital, Sichuan University, Chengdu 610041, China*
[i]*Biostatistics Unit, St Joseph's Healthcare - Hamilton, 50 Charlton Street East, Hamilton, Ontario L8N 4A6, Canada*
[j]*Department of Medicine, McMaster University, 1200 Main Street West, Hamilton, Ontario L8S 4L8, Canada*

Accepted 20 May 2019; Published online 24 May 2019

Abstract

Objective: The objective of the study was to systematically survey the methodological literature and collect suggested criteria for assessing the credibility of effect modification and associated rationales.

Study Design and Setting: We searched MEDLINE, Embase, and WorldCat up to March 2018 for publications providing guidance for assessing the credibility of effect modification identified in randomized trials or meta-analyses. Teams of two investigators independently identified eligible publications and extracted credibility criteria and authors' rationale, reaching consensus through discussion. We created a taxonomy of criteria that we iteratively refined during data abstraction.

Results: We identified 150 eligible publications that provided 36 criteria and associated rationales. Frequent criteria included significant test for interaction (n = 54), a priori hypothesis (n = 49), providing a causal explanation (n = 47), accounting for multiplicity (n = 45), testing a small number of effect modifiers (n = 38), and prespecification of analytic details (n = 39). For some criteria, we found more than one rationale; some criteria were connected through a common rationale. For some criteria, experts disagreed regarding their suitability (e.g., added value of stratified randomization; trustworthiness of biologic rationales).

Conclusion: Methodologists have expended substantial intellectual energy providing criteria for critical appraisal of apparent effect modification. Our survey highlights popular criteria, expert agreement and disagreement, and where more work is needed, including testing criteria in practice.

Keywords: Epidemiologic methods (MeSH); Meta-analysis as topic (MeSH); Clinical trials as topic (MeSH); Health care evaluation mechanisms (MeSH); Precision medicine (MeSH); Subgroup analysis

1. Introduction

Most large randomized controlled trials (RCTs) and many meta-analyses include analysis of effect modification (i.e., investigation of whether the effect of an intervention varies depending on patient characteristics such as age or intervention characteristics such as dose). Identification of true effect modification—often referred to as subgroup effect or interaction—is important for optimizing treatment for individual patients. Apparent effect modification may, however, be spurious, and if acted on may be to the patient's detriment [1].

The methodological literature has widely acknowledged the challenges of dealing with putative effect modification.

Conflict of interest: None of the authors has a conflict of interest.

* Corresponding author. Tel.: +1 289 689 1400.
E-mail address: s.schandelmaier@gmail.com (S. Schandelmaier).

https://doi.org/10.1016/j.jclinepi.2019.05.014
0895-4356/

What is new?

Key findings

- We identified 36 criteria for assessing the credibility of apparent effect modifiers and associated rationales.

What this adds to what was known?

- Key differences to previous systematic surveys include explicit definitions and eligibility criteria, a comprehensive search, rigorous methods for data collection and qualitative synthesis, and inclusion of reported rationales for suggested credibility criteria.

What is the implication and what should change now?

- The systematic survey informs those considering making subgroup claims, or evaluating subgroup claims made by others, of the most important criteria and their rationale. Further work should determine the most useful criteria and how they should be structured and implemented to develop and review subgroup claims.

In response, many methodologists have provided criteria for judging the credibility of effect modification, in particular for RCT and meta-analyses of RCTs. Examples include presence of an a priori hypothesis, use of an interaction test, and adjustment for multiplicity.

Previous groups have systematically surveyed the methodological literature addressing effect modification [2–6]. Two groups focused on credibility criteria but had important limitations in their search strategy (identifying 18 or fewer relevant publications) and methods for data abstraction [2,6]. Three other groups used more rigorous methods to survey the literature but did not explicitly focus on credibility criteria [4,5,7]; one of these had a selective focus on systematic reviews [3,4]. Moreover, these previous surveys failed to systematically collect authors' rationales for their suggested criteria.

We therefore performed a new systematic survey of the methodological literature addressing effect modification in RCTs and meta-analyses to identify credibility criteria and their associated rationales.

2. Methods

2.1. Eligibility criteria

We included publications (journal articles, reports, textbook chapters) that met the following criteria:

1 The publication devoted at least one paragraph to the interpretation of apparent effect modification (synonyms included interaction, subgroup effect, subset effect, moderation, heterogeneity of treatment effect, and predictive factor). We considered any definition of effect modification, that is, independent of the statistical approach, effect measure, or causality.
2 The publication reported one or more criteria for the credibility of apparent effect modification. We defined a criterion as a statement that links a characteristic of an apparent effect modification with an increase or decrease in credibility. We defined credibility as the extent to which an apparent effect modification represents an accurate characterization of a true underlying effect modification rather than being the result of chance or bias. We considered any paraphrase or synonym for credible including valid, true, proper, or reliable.
3 The publications addressed effect modification observed in RCTs or meta-analyses of RCTs.
4 The criteria reflect the authors' own views. We did not consider a publication if there were explicit statements that the study was a "review" of other studies, or if without that statement it obviously summarized other literature (e.g., teaching material).

2.2. Search strategy

In collaboration with an experienced medical librarian (N.B.), we developed a search strategy (Appendix A) for MEDLINE and Embase designed to capture 10 key publications of which we were already aware and which would cover a wide range of synonyms for effect medication. In addition, we applied an adapted search strategy to the WorldCat library to identify potentially eligible textbook chapters by searching their table of contents. We performed the last update in March 2018.

Teams of two methodologically trained reviewers independently screened abstracts and acquired full texts for articles that at least one reviewer deemed potentially eligible. Teams of two investigators independently assessed full texts and textbook chapters for final eligibility, resolving disagreements by discussion. One reviewer (St.S.) screened the reference lists of all eligible publications and other methodological surveys for additional potentially relevant articles and included them in the full text assessment process.

2.3. Data abstraction

We designed and pretested an online spreadsheet offering detailed instructions for data abstraction that we updated to capture issues requiring clarification as they arose. We developed a taxonomy summarizing the criteria and associated rationales as they emerged (the taxonomy provided the basis for the qualitative synthesis, see in the following). After participating in an initial calibration exercise, pairs of reviewers independently abstracted data, resolving disagreements through discussion. To ensure consistency of judgments,

St.S. was a member of all reviewer pairs (i.e., St.S. and one of Y.C., N.D., J.K., L.E.C., Y.L., A.A., T.D., and Y.Z.)

In teams of two, reviewers abstracted reported credibility and rationales offered as explanations why a criterion would increase or decrease credibility. By copying statements from the eligible articles into our data extraction forms, we captured the views of the authors verbatim. In addition to criteria and rationales, we recorded the type of publication, focus on a specific study design if any, and whether the article included a supporting simulation. When a publication provided an explicit checklist with key considerations for analysis of effect modification, we abstracted characteristics of this checklist: Number of items (not necessarily fulfilling our definition of criteria), intended audience (i.e., for users who are interpreting an apparent effect modification—which is the perspective we are taking here—or investigators who are planning an analysis of effect modification), intended study design (RCT or meta-analysis or both), presence or absence of explicit response options for item, provision of an overall judgment, and whether the development of the checklists was informed by 1) a systematic survey of the literature, 2) a formal consensus study among experts, 3) user testing (i.e., a study to find out whether users find the checklist useful and easy-to-use), 4) a reliability study (i.e., a study to find out whether users consistently rate studies in a similar way with respect to adherence to criteria), or 5) other formal methods for instrument development or testing.

2.4. Qualitative synthesis

In parallel to the data abstraction process, we developed a separate list of criteria and rationales using our own language (a taxonomy). We created new categories as they emerged. For each criterion, the taxonomy provided a keyword (e.g., a priori hypothesis), and a collection of common terms used to convey the same or related ideas (e.g., pre hoc, post hoc, exploratory, confirmatory). Reviewers used the taxonomy to organize the quotations they extracted by assigning the most closely related keywords. The taxonomy evolved in parallel with the data extraction in that reviewers could suggest new keywords when a quotation did not fit existing ones. The continuous updating of the taxonomy provided a method to involve all reviewers in the qualitative synthesis process while they were reading the publications. After completion of the data extraction, reviewers reviewed the taxonomy and suggested improvements to the wording. For each criterion and rationale, we referenced the publications from which they were extracted.

3. Results

3.1. Search results

We screened 2,117 records or journal publications and tables of contents of 151 textbooks, assessed 557 publications in full text, and finally included 150 publications (Fig. 1). The dates of publication spanned 4 decades. Publications were mostly journal articles ($n = 130$), focused on individual RCTs ($n = 97$), and provided up to 15 criteria (Table 1).

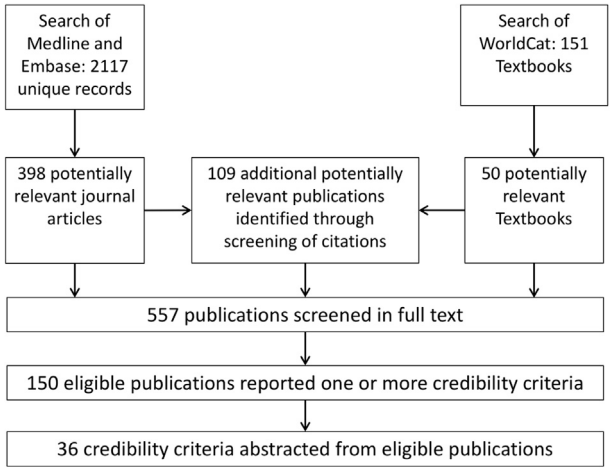

Fig. 1. Study selection flow chart.

Table 1. Characteristics of the 150 included publications

Characteristics	Frequency (total = 150)
Decade of publication	
2010s (up to 2017)	62
2000s	57
1990s	20
1980s	9
Type of publication	
Journal article	130
Textbook	16
Guidance from organization[a]	4[a]
Focus regarding study design	
RCT	97
Meta-analysis of RCTs[b]	30[b]
Both RCT and meta-analysis of RCTs	5
Explicitly any design	3
No explicit focus	15
Number of credibility criteria (according to our definition)	
1–3	77
4–6	34
7–9	18
10–12	14
13–15	7
Provides explicit checklist or instrument (see Table 3 for details)	29

Abbreviations: RCT, randomized controlled.

[a] Cochrane collaboration (REF), Center for Reviews and Dissemination (REF), European Medicines Agency (REF), The National Institute for Health and Care Excellence (REF).

[b] Includes 18 publications focusing on aggregate data meta-analysis, 4 on individual patient data meta-analysis IPD, and 8 explicitly on both.

Table 2. List of identified criteria (according to our definition; see Appendix B for reported rationales, caveats, supporting simulation studies, and references)

Category	Criterion	*n*
Design characteristics	1. Prespecification of analytic details[a]: Credibility is higher if analytic details such as cut points, time points, or statistical methods have been prespecified prior to the analysis	39
	2. Small number of candidate effect modifiers: Credibility is higher if only a small number of effect modifiers have been tested	38
	3. Within vs. between-study comparison:[b] Credibility is higher if inferences regarding effect modification are based on within-study analyses rather than a comparison of main effects across studies (individual vs. study level, individual patient data vs. aggregate data)	25
	4. Baseline characteristic vs. characteristic measured after randomization: Credibility is higher if the effect modifier is a characteristic measured at randomization, lower if the effect modifier was measured after randomization and could have been influenced by the intervention	24
	5. Power: Credibility increases with the power to detect the effect modifier	23
	6. Stratified randomization: Credibility is higher if the effect modifier was a stratification variable at randomization	15
	7. Effect modifier defined by outcome in control group:[b] Credibility is lower if, in a meta-analysis, the effect modifier is defined by outcome in the control group (e.g., ''baseline risk'' defined by mortality in control group)	6
	8. Primary outcome: Credibility is higher for effect modifiers claimed for primary rather than secondary outcomes	4
Sample characteristics	9. Sample size per subgroup: Credibility increases with sample size per subgroup and balance of sample size across subgroups	13
	10. Prognostic balance within subgroups: Credibility is higher if, within each subgroup, prognostic factors are balanced	7
	11. Measurement error: Credibility is higher if the effect modifier is measured without error, for example, no misclassification	12
Analysis characteristics	12. Interaction test: Credibility is higher if an interaction test suggests a small likelihood for a chance finding (rather than compatibility with chance or not interaction test at all) (test of homogeneity, test of heterogeneity)	54
	13. Multiplicity addressed: Credibility is higher if investigators accounted formally or informally for multiplicity (data-dredging, data-mining)	45
	14. Effect modification persists after adjusting for other effect modifiers: Credibility is higher if a multivariable analysis suggests that the apparent effect modifier is independent of other effect modifiers (confounding addressed, adjusted, joint effect vs. marginal effect, marker, proxy, surrogate)	17
	15. Assumed relationship between modifier and effect justified:[c] Credibility is higher if, for continuous effect modifiers, the researchers justified the type of relationship between effect modifier and effect, for example, linear or logarithmic	12
	16. Continuous rather than categorized:[c] Credibility is higher if continuous effect modifiers are not categorized but analyzed as a continuum	10
	17. Justified cut point:[c] If continuous effect modifiers are categorized, credibility is higher if the researchers justify the threshold, ideally a priori	1
	18. Random effects model:[b] Credibility is higher if the analysis accounts for true variation between studies (within subgroups) by applying a random effects model	9
	19. Appropriate scale: Credibility is higher if the effect modification was analyzed on an ''appropriate'' scale (with differing definitions of appropriate)	7
	20. Sensitivity analysis: Credibility is higher if a sensitivity analysis suggests robustness to relevant assumptions such as cut points or type of model	5
	21. Bayesian analysis: Credibility is higher if priors were explicitly specified and incorporated using Bayesian methods (vs. informal account of prior knowledge)	3
	22. Shrinkage applied: Credibility is higher if investigators applied shrinkage methods (weighted average of overall effect and subgroup-specific effect)	2
Numerical results	23. Quantitative vs. qualitative: Credibility is higher if the effect modification is quantitative (direction of effect consistent across levels of effect modifier) rather than qualitative (direction varies by levels of effect modifier)	12
	24. Large effect: Credibility is higher if the effect modification is large	6

(*Continued*)

Table 2. Continued

Category	Criterion	*n*
	25. Dose-response:[c] Credibility is higher if there is a dose-response relationship across ordered levels of an effect modifier	2
Contextual considerations	26. A priori hypothesis: Credibility is higher if investigators stated a hypothesis prior to performing the study, lower if an explanation arose post hoc (confirmatory vs. exploratory; hypothesis testing vs. hypothesis generating)	49
	27. Causal rationale: Credibility is higher if there is a compelling causal rationale explaining the effect modification, ideally specified a priori, and lower if not (biologic rationale, clinical rationale, other mechanism)	47
	28. Prior probability: Credibility increases with the prior probability of the effect modification being true (prior knowledge, strength of hypothesis)	16
	29. Prespecified direction: Credibility is higher if investigators correctly anticipated the direction of the subgroup effect, lower if they failed to anticipate a direction or anticipated the other direction (specific vs. vague hypothesis)	14
	30. Expert input: Credibility is higher if content expert were involved in the selection of candidate effect modifiers	5
	31. Consistent across studies: Credibility is higher if the effect modification is consistent across independent studies (reproducibility, replicability)	30
	32. Indirect evidence: Credibility is higher if indirect evidence supports the effect modifier, for example, evidence from animal studies, observational studies, related populations, or related interventions (as opposed to theory only or replication in same type of study)	20
	33. Consistent across outcomes: Credibility is higher if the effect modification is consistent across related outcomes	7
	34. Overall effect is significant: Credibility is higher if the overall effect is statistically significant (“positive” vs. “negative” trial)	14
	35. Overall effect is at low risk of bias: Credibility is lower if the overall treatment effect is at low risk of bias	7
Transparency	36. Complete reporting: Credibility is higher if all performed analyses and results are reported, ideally verified in protocol (vs. incomplete or selective reporting)	28

[a] We did not count another 26 publications that reported prespecification as a credibility criterion but without further specification (prespecification could refer to the effect modifier, the outcome, the cut point, the model, the test, or the hypothesis).
[b] Applies to meta-analysis only.
[c] Applies to continuous or ordinal effect modifiers only.

3.2. Credibility criteria

With respect to the taxonomy, we observed a saturation effect after approximately 50 publications; that is, the taxonomy changed only slightly when we abstracted the remaining 100 abstractions. Our final taxonomy included a total of 36 criteria suggested to inform the credibility of putative effect modification (Table 2). We grouped the criteria into six categories: design characteristics (8 criteria), sample characteristics (3 criteria), analysis characteristics (11 criteria), numerical results (3 criteria), contextual considerations (10 criteria), and transparency (1 criterion).

The four most frequently mentioned criteria were significant test of interaction rather than nonsignificant or no test ($n = 54$); hypothesized a priori rather than post hoc explanation ($n = 49$); strong causal (e.g., biologic) rationale rather than weak rationale or no rationale ($n = 47$); and account of multiplicity rather than ignoring multiplicity ($n = 45$). Most criteria applied to both individual RCTs and meta-analyses of RCTs. Two criteria were specific to meta-analysis: analysis of effect modification based on within rather than between-study comparisons ($n = 25$), and random rather than fixed effects model for between-study differences ($n = 9$) (Table 2).

Appendix B provides reported rationales and caveats for each criterion. For some criteria, we identified up to 4 rationales (e.g., explaining why within-study analyses are more credible than between-study analyses). Some criteria were connected through a common rationale (e.g., those addressing multiplicity). For some criteria, we did not identify an explicit rationale (e.g., why a large effect modification would be more credible than a small effect modification).

Some criteria were contentious, as suggested by conflicting rationales and caveats. For instance, a strong causal explanation (mostly framed as biologic rationale) is among the most popular criteria. Some authors have, however, argued that deducing a causal hypothesis is almost always possible and the criterion may therefore add little credibility. Some have argued that the causal explanation criterion is useful only when a causal hypothesis is absent, in which case, the credibility of the putative effect markedly diminishes. Others have argued that considerations of causality are largely irrelevant if the aim of the analysis is to identify target subgroups [8]. Other contentious criteria include whether effect modifiers are more credible when used as a stratification factor at randomization; whether qualitative effect modification is more or less credible than quantitative effect modification; or whether or not a

significant main effect increases the plausibility of an apparent effect modification (see Appendix B for details).

For 12 criteria, we found one or more supporting simulation studies (e.g., demonstrating that a formal test of interaction is more appropriate than subgroup-specific tests [9]; Appendix B).

3.3. Checklists

Thirty publications provided key considerations for analyses of effect modification in the form of explicit checklists (Table 3). The number of items per list ranged from 3 to 21 (not all of the items met our definition of credibility criteria). Fifteen checklists, varying from 3 to 16 items, were explicitly designed for users of evidence (e.g., developers of clinical practice guidelines who are critically appraising claims of effect modification). Of those, two were based on a systematic survey of the literature followed by a consensus study. None of the checklists have undergone user or reliability testing (Table 3).

4. Discussion

Many methodologists have suggested criteria for assessing the credibility of effect modification: We identified a total of 36 criteria, most of which are relevant for both individual RCTs and meta-analyses investigating effect modification (Table 2). Authors suggested some criteria—for instance tests of interaction, a priori hypotheses, or causal rationale—much more frequently than others—for example, expert input, consistency across outcomes, or overall risk of bias. For most criteria, authors provided a rationale for their choice, sometimes including caveats or reservations (Appendix B). Fifteen publications provided criteria in the form of a checklist explicitly designed for critical appraisal of apparent effect modification.

Key credibility criteria that were broadly acknowledged and well justified included the presence of a strong a priori hypothesis; analysis confined to a small number of effect modifiers; putative effect modifier is a baseline characteristic (as opposed to a characteristic observed after providing an intervention); prespecified details of the analysis of effect modification (e.g., variable definition, statistical model, time points); effect modification supported by a test of interaction; potential multiplicity taken into account; replication of the apparent effect modification across independent studies; and transparent reporting of all analyses of effect modification. A key criterion specific to meta-analysis was increased credibility if the effect modification was identified within studies (e.g., individual participant data meta-analysis) rather than by comparing summary effects between studies (e.g., meta-regression).

The identified criteria reflect two common themes in the literature regarding effect modification: one is providing safeguards against random error both on the design level (e.g., limiting number of effect modifiers and prespecifying analytic details) and on the analysis level (e.g., applying a formal test of interaction, accounting for multiplicity, shrinking estimates toward the overall, or performing a sensitivity analysis). Another common theme is consideration of external knowledge when interpreting the results (e.g., presence of a causal rationale, a priori hypothesis, indirect evidence, and replication across studies). Many methodologists noted a low confidence in claims of effect modification based on a single, typically underpowered study, and stressed external knowledge as a safeguard against spurious inferences.

Many criteria address general principles of observational research that are not specific to effect modification. For example, the most frequent criterion was whether the apparent effect modification was supported by an appropriate statistical test. Other examples include whether investigators considered confounding, prespecified analytic details, or reported all analyses. The limited attention to credibility provided in most current reports of putative subgroup effects in RCTs and meta-analyses may explain why most criteria are rather general [10,21,23,25–28,34–49].

A strength of our systematic survey is the comprehensive search. Previous systematic reviews abstracted credibility criteria from a maximum of 18 publications [2,6]; we abstracted criteria from 150 publications. We applied transparent eligibility criteria and rigorous methods for systematic data abstraction, developed a flexible taxonomy to synthesize and calibrate the views of the involved reviewers while they were abstracting the criteria, and observed a saturation effect (i.e., very few new criteria) after abstraction of approximately 50 eligible publications. We are therefore confident that we did not miss any key criteria. Another strength is that we systematically abstracted the rationales and caveats that authors offered for their criteria (Appendix B).

Our survey has limitations. The process of synthesizing verbatim quotes to characterize rationales introduced subjectivity, as did the decisions regarding lumping and splitting in the labeling of criteria. For example, a number of criteria addressed corroboration through external knowledge and we labeled these items as a priori hypothesis, causal rationale, expert input, correct anticipation of direction, prior probability, Bayesian analysis, indirect evidence, consistent across outcomes, and consistent across studies. Others may have merged these into fewer items.

Our approach may have missed certain methodological aspects that are not typically framed as credibility criteria. For instance, different methods are available to adjust for multiplicity [50], performing exploratory subgroup analyses [51], modeling continuous effect modifiers [52–54], or addressing the correlation between subgrouping variables [55]. Those considerations, however, are complex, require statistical expertise, and may therefore be impossible to frame as universal criteria.

Our findings suggest a number of inferences. The plethora of available articles addressing subgroup credibility may leave both authors of RCTs and meta-analyses, and clinicians and policy makers using their results, confused and uncertain. Most of the 15 available checklists for critical appraisal have not been developed as practical instruments. The two

Table 3. Characteristics of published checklists/instruments for effect modification

Study	Intended audience (users or investigators)[a]	Number of items[c]	Target studies	Response options for items	Informed by systematic review/consensus study?[d]
VanHoorn 2017 [6]	users	11	observational, RCT, and MA	yes, no, don't know, not applicable[e]	systematic survey and consensus study
Donegan 2015 [10]	users and investigators[b]	20	MA	yes, no	no
Burke 2015 [11]	n.s.	3	RCT	n.s.	no
European Medicines Agency 2014 [12]	users	4–5[f]	RCT	yes/no[e,f]	no
Koch 2014 [13]	n.s.	7	RCT	n.s.	no
Wang 2014a [14]	users	11	RCT	implicitly yes/no	no
Desai 2014 [15]	investigators	17	RCT	n.s.	no
Sun 2014 [16]	users	5	RCT and MA	implicitly yes/no	no
Gagnier 2013 [7]	investigators	13	MA	n.s.	systematic survey and consensus study
Varadhan 2012 [5]	investigators	9	ns	n.s.	systematic survey
Paget 2011 [17]	investigators	7	RCT	n.s.	no
Pincus 2011 [2]	users	5	RCT	implicitly yes/no	systematic survey and consensus study
Cochrane Handbook 2011 [18]	n.s.	5	MA	implicitly yes/no	no
Sun 2010 [19]	users	4	RCT and MA	implicitly yes/no	no
Kent 2010a [20]	investigators	5	RCT	n.s.	no
Fernandez 2010 [21]	users and investigators[g]	14[g]	RCT	n.s.	no
Dijkman 2009 [22]	users	16	RCT	implicitly yes/no	no
Wang 2007 [23]	investigators[h]	6	RCT	n.s.	no
Fletcher 2007 [24]	users	3	RCT	implicitly yes/no	no
Aulakh 2007 [25]	users	6	RCT	n.s.	no
Koopman 2007 [26]	investigators	6	individual participant data MA	n.s.	no
Hernandez 2006 [27]	investigators	6	RCT	n.s.	no
Bhandari 2006 [28]	users	11	RCT	implicitly yes/no	no
Rothwell 2005 [1]	n.s.	21	RCT	n.s.	no
Grouin 2005 [29]	n.s.	11	RCT	n.s.	no
Cook 2004 [30]	users	12	RCT	implicitly yes/no	no
Moreira 2002 [31]	users	8	RCT	n.s.	no
Brookes 2001 [9]	users and investigators[i]	15[i]	RCT	n.s.	no
Oxman 1992 [32]	users	7	RCT and MA	implicitly yes/no	no
Yusuf 1991 [33]	investigators	15	RCT	n.s.	no

Abbreviations: MA, meta-analysis; n.s., not specified; RCT, randomized controlled.

[a] Users refers to clinicians, systematic reviewers, guideline developers, journal editors, policy makers and other who are considering the credibility of claimed effect modification. Investigators refers to trialists or meta-analysts who are looking for guidance on how to design, carry out, or interpret their own analysis of effect modification.

[b] Criteria proposed for reporting and conduct of analyses; wording seems most appropriate for critical appraisal.

[c] We counted the items as formatted (e.g., number of list icons or rows in a table) and irrespective of our own definition for credibility criteria.

[d] We considered the following categories: systematic survey of methodological literature; formal consensus study; user testing; and reliability study. None of the studies performed user tests or test of reliability (if the purpose was critical appraisal).

[e] Includes overall judgment.

[f] Presents criteria as an algorithm with yes/no decision nodes and a final classification into credible, possibly credible, and not credible. The number of criteria depends on the path chosen.

[g] 7 items for users, 5 for investigators, 2 for editors.

[h] Reporting guideline.

[i] 11 items for investigators, 4 for users.

checklists that provide explicit response options and an overall rating have important limitations: One is a preliminary algorithm suggested by the European Medical Agency that lacks any explanation [12]. The other checklist, developed based on a systematic survey and a Delphi consensus study, addresses both prognostic factors and effect modifiers and combines credibility with applicability and clinical relevance [6]. Moreover, none of the existing checklists have been tested for feasibility, acceptability, or reliability.

Our systematic survey of reported criteria may serve as a starting point for further development of the criteria-based approach credibility of effect modification. A formal instrument that overcomes the limitations of previous credibility checklists is urgently needed. About 20% of published trials and meta-analyses include claims of effect modification [10,23,27,39,40,42,43], and difficulty in determining their credibility frequently compromises guideline development and decision-making. Currently, those considering the credibility of subgroup analyses have a profusion of alternative checklists from which to choose with, because of their limitations, none that are clearly superior to the others.

Criteria that are widely acknowledged or strongly supported by simulation studies could provide the basis for a new instrument. Criteria for which we identified conflicting or missing rationales or caveats seem less suitable or would require modification. Development of a new instrument would require careful attention to the target audience (e.g., trial investigators or systematic review authors who are considering to claim an effect modification, or clinicians, guideline developers, journal editors, or policy makers who are evaluating a claim of effect modification). Determining the feasibility, acceptability, and reliability of any instrument suggested for wide use would therefore be crucial.

CRediT authorship contribution statement

Stefan Schandelmaier: Conceptualization, Data curation, Investigation, Formal analysis, Funding acquisition, Methodology, Project administration, Writing - original draft, Writing - review & editing. **Yaping Chang:** Investigation, Formal analysis, Writing - review & editing. **Nivedita Devasenapathy:** Investigation, Formal analysis, Writing - review & editing. **Tahira Devji:** Investigation, Formal analysis, Writing - review & editing. **Joey S.W. Kwong:** Investigation, Formal analysis, Writing - review & editing. **Luis E. Colunga Lozano:** Investigation, Formal analysis, Writing - review & editing. **Yung Lee:** Investigation, Formal analysis, Writing - review & editing. **Arnav Agarwal:** Investigation, Formal analysis, Writing - review & editing. **Neera Bhatnagar:** Conceptualization, Methodology, Writing - review & editing. **Hannah Ewald:** Methodology, Investigation, Formal analysis, Writing - review & editing. **Ying Zhang:** Investigation, Formal analysis, Writing - review & editing. **Xin Sun:** Conceptualization, Methodology, Writing - review & editing. **Lehana Thabane:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Michael Walsh:** Conceptualization, Methodology, Supervision, Writing - review & editing. **Matthias Briel:** Conceptualization, Investigation, Methodology, Supervision, Writing - review & editing. **Gordon H. Guyatt:** Conceptualization, Investigation, Formal analysis, Methodology, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgments

The authors thank Kuebra Oezoglu for gathering full-text articles.

This work was supported by the Swiss National Science Foundation [grant number P300PB_164750]; the Gottfried and Julia Bangerter-Rhyner-Foundation; and the Freiwillige Akademische Gesellschaft Basel. The funders were not involved in the study design; collection, analysis or interpretation of the data; writing of the report; or decision to submit the article for publication.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2019.05.014.

References

[1] Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86.

[2] Pincus T, Miles C, Froud R, Underwood M, Carnes D, Taylor SJ. Methodological criteria for the assessment of moderators in systematic reviews of randomised controlled trials: a consensus study. BMC Med Res Methodol 2011;11:14.

[3] Gagnier JJ, Moher D, Boon H, Beyene J, Bombardier C. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. BMC Med Res Methodol 2012;12:111.

[4] West SL, Gartlehner G, Mansfield AJ, Poole C, Tant E, Lenfestey N, et al. Comparative Effectiveness Review Methods: Clinical Heterogeneity. Agency for Healthcare Research and Quality; 2010: Methods Research Paper AHRQ Publication No 10-EHC070-EF. Available at http://effectivehealthcareahrqgov/. Accessed March 1, 2019.

[5] Varadhan R, Stuart EA, Louis TA, Segal JB, Weiss CO. Review of guidance documents for selected methods in patient centered outcomes research: standards in addressing heterogeneity of treatment effectiveness in observational and experimental patient centered outcomes research 2012. Available at pcori.org. Accessed March 1, 2019.

[6] van Hoorn R, Tummers M, Booth A, Gerhardus A, Rehfuess E, Hind D, et al. The development of CHAMP: a checklist for the appraisal of moderators and predictors. BMC Med Res Methodol 2017;17:173.

[7] Gagnier JJ, Morgenstern H, Altman DG, Berlin J, Chang S, McCulloch P, et al. Consensus-based recommendations for investigating clinical heterogeneity in systematic reviews. BMC Med Res Methodol 2013;13:106.

[8] VanderWeele TJ, Knol MJ. Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. Ann Intern Med 2011;154:680–3.

[9] Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: quantifying the risks of false-positives and false-negatives. Health Technol Assess 2001;5(33):1–56.

[10] Donegan S, Williams L, Dias S, Tudur-Smith C, Welton N. Exploring treatment by covariate interactions using subgroup analysis and meta-

regression in cochrane reviews: a review of recent practice. PLoS One 2015;10:e0128804.
[11] Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. BMJ 2015;351:h5651.
[12] European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials (draft). London, UK: European Medicines Agency; 2014.
[13] Koch A, Framke T. Reliably basing conclusions on subgroups of randomized clinical trials. J Biopharm Stat 2014;24(1):42−57.
[14] Wang SJ, Hung HM. A regulatory perspective on essential considerations in design and analysis of subgroups when correctly classified. J Biopharm Stat 2014;24(1):19−41.
[15] Desai M, Pieper KS, Mahaffey K. Challenges and solutions to pre- and post-randomization subgroup analyses. Curr Cardiol Rep 2014; 16(10):531.
[16] Sun X, Ioannidis JP, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. JAMA 2014;311:405−11.
[17] Paget MA, Chuang-Stein C, Fletcher C, Reid C. Subgroup analyses of clinical effectiveness to support health technology assessments. Pharm Stat 2011;10(6):532−8.
[18] Higgins JP, Green SB, editors. Cochrane handbook for systematic reviews of interventions Version 5.1.0. The Cochrane Collaboration; 2011.
[19] Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010;340:c117.
[20] Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials 2010;11:85.
[21] Fernandez YGE, Nguyen H, Duan N, Gabler NB, Kravitz RL. Assessing heterogeneity of treatment effects: are authors misinterpreting their results? Health Serv Res 2010;45:283−301.
[22] Dijkman B, Kooistra B, Bhandari M, Evidence-Based Surgery Working Group. How to work with a subgroup analysis. Can J Surg 2009; 52(6):515−22.
[23] Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine–reporting of subgroup analyses in clinical trials. N Engl J Med 2007;357:2189−94.
[24] Fletcher J. Subgroup analyses: how to avoid being misled. BMJ 2007; 335:96−7.
[25] Aulakh AK, Anand SS. Sex and gender subgroup analyses of randomized trials. Womens Health Issues 2007;17(6):342−50.
[26] Koopman L, van der Heijden GJ, Glasziou PP, Grobbee DE, Rovers MM. A systematic review of analytical methods used to study subgroups in (individual patient data) meta-analyses. J Clin Epidemiol 2007;60:1002−9.
[27] Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading? Am Heart J 2006;151:257−64.
[28] Bhandari M, Devereaux PJ, Li P, Mah D, Lim K, Schunemann HJ, et al. Misuse of baseline comparison tests and subgroup analyses in surgical trials. Clin Orthop Relat Res 2006;447:247−51.
[29] Grouin JM, Coste M, Lewis J. Subgroup analyses in randomized clinical trials: statistical and regulatory issues. J Biopharm Stat 2005; 15(5):869−82.
[30] Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. Med J Aust 2004;180(6):289−91.
[31] Moreira ED, Susser E. Guidelines on how to assess the validity of results presented in subgroup analysis of clinical trials. Rev Hosp Clin Fac Med Sao Paulo 2002;57(2):83−8.
[32] Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992;116:78−84.
[33] Yusuf S, Wittes J, Probstfield J, Tyroler HA. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 1991;266:93−8.
[34] Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. JAMA Intern Med 2017;177(4):554−60.
[35] Wallach JD, Sullivan PG, Trepanowski JF, Steyerberg EW, Ioannidis JP. Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses. BMJ 2016;355:i5826.
[36] Gabler NB, Duan N, Liao D, Elmore JG, Ganiats TG, Kravitz RL. Dealing with heterogeneity of treatment effects: is the literature up to the challenge? Trials 2009;10:43.
[37] Saragiotto BT, Maher CG, Moseley AM, Yamato TP, Koes BW, Sun X, et al. A systematic review reveals that the credibility of subgroup claims in low back pain trials was low. J Clin Epidemiol 2016;79:3−9.
[38] 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.
[39] Zhang S, Liang F, Li W, Hu X. Subgroup analyses in reporting of phase III clinical trials in solid tumors. J Clin Oncol 2015;33(15): 1697−702.
[40] Barton SP C, Sclafani F, Cunningham D, Chau I. 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(18):2732−9.
[41] Mistry D, Patel S, Hee SW, Stallard N, Underwood M. Evaluating the quality of subgroup analyses in randomized controlled trials of therapist-delivered interventions for nonspecific low back pain: a systematic review. Spine (Phila Pa 1976) 2014;39(7):618−29.
[42] Kasenda B, Schandelmaier S, Sun X, von Elm E, You J, Blumle A, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. BMJ 2014;349:g4539.
[43] Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ 2012;344:e1553.
[44] Koopman L, van der Heijden GJ, Hoes AW, Grobbee DE, Rovers MM. Empirical comparison of subgroup effects in conventional and individual patient data meta-analyses. Int J Technol Assess Health Care 2008;24(3):358−61.
[45] Patsopoulos NA, Tatsioni A, Ioannidis JP. Claims of sex differences: an empirical assessment in genetic associations. JAMA 2007;298: 880−93.
[46] Higgins JT S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. J Health Serv Res Policy 2002;7:51−61.
[47] Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Stat Med 2002;21:2917−30.
[48] Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000;355: 1064−9.
[49] Parker AB, Naylor CD. Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials. Am Heart J 2000;139:952−61.
[50] Alosh M, Huque MF. Multiplicity considerations for subgroup analysis subject to consistency constraint. Biom J 2013;55(3):444−62.
[51] Lipkovich I, Dmitrienko A, Agostino BRD' Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med 2017;36:136−96.
[52] 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.
[53] 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.
[54] 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.
[55] Varadhan R, Wang SJ. Standardization for subgroup analysis in randomized controlled trials. J Biopharm Stat 2014;24(1):154−67.