EVALUATING THE CREDIBILITY OF EFFECT MODIFICATION CLAIMS IN RANDOMIZED CONTROLLED TRIALS AND META-ANALYSES

By STEFAN SCHANDELMAIER, MMus, MD, MSc

A Thesis Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

McMaster University (Health Research Methodology), Hamilton, Ontario

DOCTOR OF PHILOSOPHY (2019)

TITLE: Evaluating the Credibility of Effect Modification Claims in Randomized Controlled Trials and Meta-Analyses

AUTHOR: Stefan Schandelmaier, MMus, MD, MSc

SUPERVISOR: Distinguished Professor Gordon Guyatt, MD, MSc, FRCP, OC

NUMBER OF PAGES: ix, 158

Lay abstract

Randomized controlled trials and meta-analyses provide the best available evidence to evaluate whether effects of a therapy vary among individual patients. Efforts to decide whether treatment effects differ across patients are important and frequently done but difficult to interpret. The fundamental challenge is to decide whether apparent differences in effect are real or due to chance. To aid this decision, experts have suggested various sets of credibility criteria, all with important limitations. This thesis documents how we systematically addressed the limitations of previous approaches. Key steps were a systematic survey of the available credibility criteria, a consensus study among leading methodologists, and a formal user-testing study. The result is a new instrument for assessing the credibility of effect modification analyses (ICEMAN).

Abstract

Background: Many randomized controlled trials (RCTs) and meta-analyses include analyses of effect modification (also known as subgroup, interaction, or moderation analyses). Methodologists have widely acknowledged the challenges in deciding whether an apparent effect modification is credible or likely the result of chance or bias. Various sets of credibility criteria are available (Chapter 2 provides an example) but are inconsistent, vague in wording, lack guidance for deciding on overall credibility, and have not been systematically tested.

Objective: To systematically develop a formal instrument to assess the credibility of effect modification analyses (ICEMAN) in RCTs and meta-analyses of RCTs.

Methods: Key steps in the development process included 1) a systematic survey of the literature to identify available criteria, rationales, and previous instruments, 2) a formal consensus study among 10 leading experts, and 3) a formal user-testing study to refine the instrument based on interviews with trial investigators, systematic reviewer authors, and journal editors who applied drafts of the instrument to published claims of effect modification.

Results: The systematic survey identified 150 relevant publications, 36 candidate credibility criteria with associated rationales, and 30 existing checklists (Chapter 3). The consensus study consisted of two main video conferences and multiple rounds of written discussion. The user-testing involved 17 users (including systematic review authors, trial investigators, and journal editors) who suggested substantial improvements based on detailed interviews. The final instrument provides separate versions for RCTs (five core questions) and meta-analyses (eight core questions) with explicit response options, and an overall credibility rating ranging from very low to high credibility. A detailed manual provides rationales, supporting references, examples from the literature, and suggestions for use in combination with other quality appraisal tools and reporting (Chapter 4).

Discussion: ICEMAN is a rigorously developed instrument to evaluate claims of effect modification and addresses the main limitations of previous approaches.

Acknowledgements

Thank you, Gordon, for being such a wonderful mentor and teacher, and exemplifying an academic style in such an efficient, sophisticated, philanthropic, and fascinating way that it will inspire me throughout my career – I will carry the flag!

Thank you, Mike, Lehana, and Matthias, for guiding me smoothly through the PhD program and providing me with insightful feedback, smart advice, and motivation whenever I needed it.

Thank you to my brilliant colleagues and coauthors for being so generous with their time and contributions that not only made this work possible but raised it to a level much higher than I could have reached without them.

Thank you to the four Swiss institutions who provided me with salary and research support to make this work and learning experience possible.

Thank you, Mama, Doris, and Paul, for your company in Canada, loving child care, support packages, and everything else.

Thank you, Anne, Mira, and Felix, for relieving me of responsibilities and your love that gave me the energy for this work and made me happy and proud every day!

Table of contents

Lists of tables, figures, and appendices

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Abbreviations

1. BMJ – British Medical Journal
2. CI – Confidence interval
3. GRADE – Grading of Recommendations Assessment, Development and Evaluation
4. ICEMAN – Instrument for assessing the credibility of effect modification analyses
5. IPD – Individual participant data
6. LIPUS – Low intensity pulsed ultrasound
7. RCT – randomized controlled trial

Declaration of Academic Achievement

Chapter 1: Unpublished. Stefan Schandelmaier is the author.

Chapter 2: Published in the British Medical Journal (BMJ 2017;356:j656). JWB, RAS, GHG and POV conceived the study idea. SS and JWB coordinated the systematic review. SS wrote the first draft of the manuscript. RC designed the search strategy. LL, AK, RC, and SS screened abstracts and full texts. LL, AK, RAS, TA, and SS acquired the data and judged risk of bias in the studies. SS, JWB and DHA performed the data analysis. DHA and GHG provided statistical advice. SS, RAS, POV, JWB and GHG interpreted the data analysis. All authors critically revised the manuscript. SS had full access to all of the data in the study, and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Chapter 3: In review at the Journal of Clinical Epidemiology. SS, GHG, LT, MW, MB, and XS conceived the study idea. SS coordinated the study. SS wrote the first draft of the manuscript. SS, NB, and HE designed the search strategy. SS, MB, and HE screened abstracts and full texts. YC, ND, TD, JK, LC, YL, AA, YZ, MB, and SS acquired the data and performed the qualitative data analysis. All authors interpreted the results and critically revised the manuscript. SS had full access to all of the data in the study, and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Chapter 4: Ready for submission to the British Medical Journal. SS, GHG, LT, MW, MB, and XS conceived the study idea. SS coordinated the study. SS wrote the first draft of the manuscript. SS, MB, RV, CS, RH, JG, MB, GV, ID, XS, WS, MW, JI, LT, and GG participated in the consensus study. SS and ND performed and transcribed the interviews. All authors interpreted the results and critically revised the manuscript. SS had full access to all of the data in the study, and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Chapter 5: Unpublished. Stefan Schandelmaier is the author.

Chapter 1: Introduction

Randomized controlled trials (RCTs) provide the optimal design to test the causal effects of health care interventions and therefore play a crucial role in evidence-based decision making.

The main goal of an RCT – or a meta-analysis if more than one RCT is available – is to quantify the effect of an intervention for the entire study population, e.g. in the form of an overall risk ratio. The overall effect best reflects the effect as it would occur in a single patient characterized by the average of all measured and unmeasured characteristics of the study population. For instance, a cardiovascular trial may test the effect of a new drug to prevent myocardial infarction in a population that includes men and women, and 30% of the participants have diabetes. Then, if the study suggests an overall benefit, e.g. a risk ratio of 0.8, the average patient to whom this effect applies would be a mixture between men and women with 30% diabetes. Of course, such a patient does not exist and the list of impossible average characteristics could be easily extended.

The example illustrates that, in order to apply overall effects to an individual patient, we have to assume that the treatment effect is consistent across characteristics in which the target patient differs from the average: Using again the example, let us assume we want to apply the overall effect to someone who could have participated in the study, e.g. a man with diabetes. To apply the overall effect, we would have to assume that it is valid for both men and women, diabetics and non-diabetics, and any other patient characteristics in which our patient might differs from the average. In other words, we assume that none of the varying patient characteristics would substantially influence the suggested treatment effect.

Such a consistency assumption may seem strong to many, in particular to health care practitioners who experience the diversity of patients in their daily practice, and also to meta-analysts who are used to see inconsistent treatment effects across studies. Therefore, it is natural that researchers scrutinize the consistency assumption and explore whether a treatment effect might vary across patient characteristics. Those analyses are called *analyses of effect modification* and the examined patient characteristics *potential effect modifiers*.

Analyses of effect modification vary with respect to terminology and methods. Most common are subgroup analyses for which investigators repeat the main analysis within subgroups (e.g. men and women) and then compare the resulting effect estimates. If they differ, the characteristic (e.g. sex) is said to be an effect modifier, moderator, a predictor of response, or that there is an interaction between sex and the intervention. If the potential effect modifier is a continuous variable, e.g. age, effect modification can also be analyzed continuously, usually by including an interaction term in a regression model. In the context of meta-analysis, it is common to compare subgroups of studies or use meta-regression for continuous effect modifiers. Analyses of effect modification become more complex if it is possible to define both subgroups of patients and subgroups of studies as in most individual participant data meta-analyses.

There are a number of hypothesis tests available to assess the extent to which an apparent effect modification is compatible with chance.[1-3] Usually, they test the null hypothesis that a treatment effect is consistent across levels of the candidate effect modifier. Other possible, though rarely applied, null hypotheses are that effects of individual subgroups have a consistent direction (tests of qualitative interaction),[4] or that effects of individual subgroups do not follow a specific pattern (tests of trends).[5] Another class of analyses are data-driven algorithms that identify subgroups of patients who show the most extreme effects [6] or mathematical functions that best describe the relationship between a continuous effect modifier and the treatment effect.[7]

Irrespective of the specific methods used, analyses of effect modification share three fundamental challenges that complicate their interpretation:

1) A large number of candidate effect modifiers: While there are usually only a small number of candidate therapies available for a condition, the number of potential effect modifiers can be very large and vary from study to study. Essentially any measured baseline characteristic could be analyzed for possible effect modification. In the context of meta-analyses, additional candidate effect modifiers are study characteristics such as study design, study quality, or type of intervention. The large number of candidate effect modifiers leads to two subsequent issues: potential multiplicity issues (multiple hypothesis tests within studies may compromise the results of hypothesis tests) and uniqueness (if a study proposes an effect modification, none of the previous and subsequent studies might consider the same effect modifier using the same methods, thus making successful replication extremely rare).[8]

2) **Error:** Analysis of effect modification are prone to random and systematic error. Reasons for the high risk of random error include multiplicity, but also low power and the low prior probability of an effect modification being true.[9] Analysis of effect modification performed in meta-analyses additionally suffer from systematic errors that have to do with aggregation of data and study-level confounding.[10]

There is clearly a gap between the available methodological knowledge and current practice as documented in numerous meta-studies.[11-24] For instance, they show that only a minority of published analyses of effect modification include a test of interaction, justify the choice of effect measure, acknowledge the risk for confounding, treat continuous effect modifiers as continuous, or, in IPD meta-analyses, describe whether the effect modifier of interest varies mainly within studies (more credible) or between studies (less credible).[11-24]

3) **Insufficient reporting**: The meta-studies also document that reporting of analyses of effect modification is frequently insufficient, both in protocols [15,17] and final reports. [11-24] Potential reasons are the secondary character of the analyses and the potentially large number of details documentation of which might be burdensome or perceived as not relevant (e.g. documentation of all considered candidate effect modifiers and definitions, hypotheses, effect measures, outcomes, and methods and results of interaction tests).

In response to the many challenges, methodologists have suggested guidance for performing, interpreting, and reporting of analyses of effect modification. This thesis addresses interpretation, more specifically the critical appraisal of a putative effect modification identified in a RCT or a meta-analysis of RCTs – for which we developed a new quality appraisal tool called Instrument for assessing the Credibility of Effect Modification ANalysis (ICEMAN).

Chapter 2 provides an example illustrating the status quo before ICEMAN was available. In a meta-analysis investigating the effects of a popular ultrasound device on bone healing, we found a potentially credibility effect modification: the risk of bias of individual RCTs seemed to be associated with the size of effect; high risk of bias studies showed a clear increase in bone healing, whereas low risk of bias studies showed essentially no effect.[25] The best available criteria we knew was a previously published checklist with 11 items. We applied the checklist and included it as a table in the final publication.[26] Based on the checklist, we interpreted the potential effect modification as credible and our assessment convinced reviewers and editors. In addition, the assessment convinced an associated guideline panel who decided to base their recommendations exclusively on the subgroup with low risk of bias trials resulting in a strong recommendation against the use of ultrasound for accelerating bone healing.[27]

Although we perceived the 11 proposed credibility criteria as useful, we also realized a number of major limitations: lack of explicit response options, overlap between items, some items were not applicable, no standardized format available for presentation, and lack of an overall rating. In addition, the criteria were not developed based on standard methods for instrument development such as a systematic search for candidate items, a formal item selection process informed by an expert panel, and user-testing.[28,29] Consequently, we decided to perform a systematic survey of the methodological literature addressing the credibility of claims of effect modification and develop a formal credibility instrument (unless the systematic survey would identify an acceptable instrument).

Chapter 3 presents the methods and results of the systematic survey. In additional to the 11 criteria with which we were familiar, we identified another 25 criteria provided in 150 journal articles and text book chapters. The survey also identified 29 other sets of criteria, all of which had some important limitations. Therefore, we decided to move on with the development process of the new instrument.

Chapter 4 documents in detail the development process that included an initial concept and, informed by the results of the systematic survey, a formal expert consensus and a user testing study. The chapter also includes the final ICEMAN instrument –separate versions for RCTs and meta-analyses of RCTs – and a discussion of the new instruments' strength and limitations.

Chapter 5 provides a discussion of the advantages of ICEMAN over previous approaches and remaining limitations, our strategy to implement ICEMAN in practice, and prospects for future research.

References

1.Greenland S. **Tests for interaction in epidemiologic studies: a review and a study of power.** Statistics in medicine. 1983;2(2):243-51.
2.Alosh M, Huque MF, Bretz F, D'Agostino RB, Sr. **Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials.** Statistics in medicine. 2017;36(8):1334-60.
3.Sedgwick P. **Randomised controlled trials: tests of interaction.** Bmj. 2014;349:g6820.
4.Kitsche A. **Detecting qualitative interactions in clinical trials with binary responses.** Pharmaceutical statistics. 2014;13(5):309-15.
5.Chemoradiotherapy for Cervical Cancer Meta-Analysis C. **Reducing uncertainties about the effects of chemoradiotherapy for cervical cancer: a systematic review and meta-analysis of individual patient data from 18 randomized trials.** Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2008;26(35):5802-12.
6.Lipkovich I, Dmitrienko A, B. R. D' Agostino S. **Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials.** Statistics in medicine. 2017;36(1):136-96.
7.Royston P, Sauerbrei W. **A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials.** Statistics in medicine. 2004;23(16):2509-25.
8.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.
9.Burke JF, Sussman JB, Kent DM, Hayward RA. **Three simple rules to ensure reasonably credible subgroup analyses.** Bmj. 2015;351:h5651.
10.Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. **Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?** Bmj. 2017;356:j573.
11.Schuit E, Li AH, Ioannidis JPA. **How often can meta-analyses of individual-level data individualize treatment? A meta-epidemiologic study.** International journal of epidemiology. 2018.
12.Gabler NB, Duan N, Raneses E, Suttner L, Ciarametaro M, Cooney E, et al. **No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals.** Trials. 2016;17(1):320.
13.Simmonds M, Stewart G, Stewart L. **A decade of individual participant data meta-analyses: A review of current practice.** Contemporary clinical trials. 2015;45(Pt A):76-83.
14.Zhang S, Liang F, Li W, Hu X. **Subgroup Analyses in Reporting of Phase III Clinical Trials in Solid Tumors.** Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2015;33(15):1697-702.
15.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(6):e0128804.
16.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.** European journal of cancer. 2015;51(18):2732-9.

17.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.
18.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.
19.Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. **Statistics in medicine--reporting of subgroup analyses in clinical trials.** The New England journal of medicine. 2007;357(21):2189-94.
20.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.** Journal of clinical epidemiology. 2007;60(10):1002-9.
21.Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. **Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading?** American heart journal. 2006;151(2):257-64.
22.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.** Clinical orthopaedics and related research. 2006;447:247-51.
23.Moreira ED, Jr.; Stein, Z.; Susser, E. **Reporting on methods of subgroup analysis in clinical trials: a survey of four scientific journals.** Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica [et al]. 2001;34(11):1441-6.
24.Assmann SF, Pocock SJ, Enos LE, Kasten LE. **Subgroup analysis and other (mis)uses of baseline data in clinical trials.** Lancet. 2000;355(9209):1064-9.
25.Schandelmaier S, Kaushal A, Lytvyn L, Heels-Ansdell D, Siemieniuk RA, Agoritsas T, et al. **Low intensity pulsed ultrasound for bone healing: systematic review of randomized controlled trials.** Bmj. 2017;356:j656.
26.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.
27.Poolman RW, Agoritsas T, Siemieniuk RA, Harris IA, Schipper IB, Mollon B, et al. **Low intensity pulsed ultrasound (LIPUS) for bone healing: a clinical practice guideline.** Bmj. 2017;356:j576.
28.Streiner DL, Norman GR, Cairney J. Health measurement scales : a practical guide to their development and use. Oxford: Oxford University Press; 2015.
29.Whiting P, Wolff R, Mallett S, Simera I, Savovic J. **A proposed framework for developing quality assessment tools.** Systematic reviews. 2017;6(1):204.

Chapter 2: Low intensity pulsed ultrasound for bone healing: systematic review of randomized controlled trials

Published in BMJ 2017;356:j656; The BMJ explicitly allows republication in a book or other publication edited by the author without asking their permission (and subject only to acknowledging first publication in The BMJ and giving a full reference or web link, as appropriate. (http://www.bmj.com/permissions)

Stefan Schandelmaier, *methodologist*[1,2]; Alka Kaushal, *physician*[1,3]; Lyubov Lytvyn, *methodologist*[4]; Diane Heels-Ansdell, *biostatistician*[1]; Reed A.C. Siemieniuk, *methodologist*[1,5]; Thomas Agoritsa s, *assistant professor*[1,6]; Gordon H Guyatt, *distinguished professor*[1,7]; Per O Vandvik, *associate professor*[8,9]; Rachel Couban, *medical librarian*[3]; Brent Mollon, *orthopedic surgeon*[10]; Jason W. Busse, *associate professor*[1,3,11]

1. Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton, Ontario, Canada
2. Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Basel, Switzerland
3. Michael G. DeGroote Institute for Pain Research and Care, McMaster University, Hamilton, Canada
4. Oslo University Hospital, Oslo, Norway
5. Department of Medicine, University of Toronto, Toronto, Ontario, Canada
6. Division General Internal Medicine & Division of Clinical Epidemiology, University Hospitals of Geneva, Geneva, Switzerland
7. Department of Medicine, McMaster University, Hamilton, Ontario, Canada
8. Institute of Health and Society, Faculty of Medicine, University of Oslo, Oslo, Norway
9. Department of Medicine, Innlandet Hospital Trust-division, Gjøvik, Norway
10. Orillia Soldiers' Memorial Hospital, Orillia, Ontario, Canada
11. Department of Anesthesia, McMaster University, Hamilton, Canada

Abstract

Objective: To determine the efficacy of low intensity pulsed ultrasound (LIPUS) for healing of fracture or osteotomy.

Design: Systematic review and meta-analysis**.**

Data sources: MEDLINE, EMBASE, CINAHL, the Cochrane Central Register of Controlled Trials, and trial registries up to November 2016.

Study selection: Randomised controlled trials (RCTs) comparing LIPUS to sham device or no device in patients with any kind of fracture or osteotomy.

Review methods: Two independent reviewers identified studies, extracted data, and assessed risk of bias. A parallel guideline committee (*BMJ* Rapid Recommendation) provided input on the design and interpretation of the systematic review, including selection of patient-important outcomes. We assessed the quality of evidence using GRADE.

Results: We included 26 RCTs with a median sample size of 30 (range 8 to 501). The most trustworthy evidence came from four trials at low risk of bias including patients with tibia or clavicle fractures. Compared with control, LIPUS does not reduce time to return to work (percent difference: 2.7% later with LIPUS, 95% confidence interval [CI] 7.7% earlier to 14.3% later, moderate certainty) or the number of subsequent operations (risk ratio: 0.80, 95% CI 0.55 to 1.16, moderate certainty). For pain, days to weight bearing, and radiographic healing, effects varied substantially between studies. For all three outcomes, trials at low risk of bias failed to demonstrate a benefit with LIPUS, while trials at high risk of bias suggested a benefit (interaction $p<0.001$). Considering only low risk of bias trials, LIPUS does not reduce days to weight bearing (4.8% later, 95% CI 4.0% earlier to 14.4 % later, high certainty), pain at 4 to 6 weeks (mean difference on 0-100 visual analogue scale: 0.94 lower, 95% CI 2.54 lower to 0.65 higher, high certainty), and days to radiographic healing (1.7% earlier, 95% CI 11.2% earlier to 8.8% later, moderate certainty).

Conclusions: Based on moderate to high quality evidence from studies in patients with fresh fracture, LIPUS does not improve patient-important outcomes and probably has no effect on radiographic bone healing.

Registration: PROSPERO CRD42016050965

What is already known of this topic
Low intensity pulsed ultrasound (LIPUS) devices are marketed worldwide to accelerate recovery from a fracture or osteotomy.

Previous systematic reviews provided no definite conclusions about the effect of LIPUS on patient-important outcomes and radiographic healing.

What this study adds
A guideline panel including patients and clinical experts informed outcome selection, importance of outcomes, subgroup analyses, and interpretation of results.

Subgroup analyses suggested that beneficial effects of LIPUS are restricted to trials at high risk of bias.

With inclusion of the recently published TRUST trial, sufficient high quality data for patients with fresh fractures has accumulated to conclude that LIPUS fails to improve patient-important outcomes and radiographic healing.

INTRODUCTION

For over 20 years, patients have used low intensity pulsed ultrasound (LIPUS) as an adjunct therapy to improve bone healing. Based on radiographic outcomes, the US Food and Drug Administration and the UK National Institute for Health and Care Excellence NICE have approved LIPUS for fracture healing.[1,2] Depending on country and device model, LIPUS devices currently cost between £1000-4000. In 2008, 45% of Canadian trauma surgeons prescribed bone stimulators to manage tibia fractures, equally split between LIPUS and electrical stimulation (21% each).[3] Sales from LIPUS amounted to approximately $250 million in 2006 in the US alone.[3,4]

Within the last seven years, 10 systematic reviews have assessed the effectiveness of LIPUS for bone healing.[5-14] Because existing randomised controlled trials (RCTs) were limited by small sample size, risk of bias, inconsistent results, and failure to address patient-important outcomes, no review offered definitive conclusions. All reviews identified the need for additional RCTs. In addition, recent reviews used suboptimal strategies for outcome selection, data synthesis analysis, and interpretation, leading to potentially misleading conclusions. For instance, the most recent systematic review, published in the top speciality journal for orthopaedic surgeons, considered radiographic union a "critically important outcome" and did not assess the effect of LIPUS on the patient-important outcomes of pain relief or re-operation. Their conclusion that "LIPUS treatment effectively reduces the time to radiographic fracture union" is questionable because it is based on the pooled absolute difference in days to healing, which does not account for the large variation in healing time, showed high unexplained heterogeneity (I^2= 94%), and was driven by studies at high risk of bias. This positive conclusion has the potential to expand the already considerable use of a potentially ineffective therapy.

This systematic review is part of the *BMJ* Rapid Recommendations project, a collaborative effort from the MAGIC research and innovation program (www.magicproject.org) and *The BMJ.* The aim of the project is to respond to new potentially practice-changing evidence and provide a trustworthy practice guideline in a timely manner.[15] In this case, the publication of the TRUST trial,[16] a multicentre trial that randomised 501 patients with tibia fractures and has cast doubt on the effectiveness of LIPUS, initiated the process. This systematic review informed a parallel guideline published in a multi-layered electronic format on *The BMJ*[17] and MAGICapp (https://www.magicapp.org/public/guideline/mL6yYj).

Our objective was to assess whether LIPUS compared to sham device or no device improves patient important outcomes and radiographic healing in patients with any kind of fracture or osteotomy.

METHODS

Guideline panel and patient involvement

According to the *BMJ* Rapid Recommendations process,[15] a guideline panel provided critical oversight to the review and identified populations, subgroups, and outcomes of interest. The panel included six content experts (five orthopaedic or trauma surgeons and one physiotherapist), six methodologists (four of whom are also front-line clinicians), and four patients with personal experience of fractures (one of whom had used LIPUS). All patients received personal training and support to optimise contributions throughout the guideline development process. The patient panel members led the interpretation of the results based on what they expected the typical patient values and preferences to be, as well as the variation between patients.

Information sources

We searched MEDLINE, PubMed, EMBASE, CINAHL, and the Cochrane Central Register of Controlled Trials up to 16 November 2016, using a combination of keywords and MeSH terms for fracture, orthopaedic surgical procedures, and ultrasound. Additional searches included trials registries clinicaltrials.gov and isrctn.com. An experienced research librarian designed the search strategies (appendix 1). Two independent reviewers scanned the references from eligible studies, related systematic reviews, and all studies citing eligible RCTs on Google Scholar.

Study selection

We included RCTs comparing LIPUS to a sham device or no device in patients with any type of fracture regardless of location (long-bone or other bone), type (fresh fracture, delayed union, non-union, or stress fracture), or clinical management (operative or non-operative). We included any type of osteotomy, including distraction osteogenesis. We excluded trials published only as protocol or abstract if attempts to get the final results from investigators were unsuccessful.

Two reviewers, independently and in duplicate, screened the titles and abstracts of identified articles and acquired the full text of any article that either reviewer judged to be potentially eligible. They independently applied the eligibility criteria to the full texts and, when consensus could not be reached, resolved disagreements through discussion or adjudication by a third reviewer.

Data collection

Two reviewers used standardised forms to independently abstract data; they resolved disagreements by discussion or involved a third reviewer when required. Extracted data included patient characteristics, fracture characteristics, clinical management, risk of bias, intervention details, statements about compliance with treatment, and outcomes.

Risk of bias assessment

Two reviewers independently assessed risk of bias using a modified Cochrane risk of bias instrument that includes response options of "definitely or probably yes" (assigned a low risk of bias) or "definitely or probably no" (assigned a high risk of bias), an approach we have previously validated.[18] On the study level, we assessed generation of randomisation sequence, concealment of allocation, blinding of patients, caregivers, and outcome reporting (by comparing each publication with their corresponding published protocol, when available). For each outcome within studies, we assessed blinding of outcome assessors, loss to follow-up, and additional limitations. We considered ≥20% loss-to follow-up to represent a high risk of bias unless the investigators performed appropriate sensitivity analyses demonstrating the robustness of the results. As a sensitivity analysis, we alternatively considered a more conservative threshold of ≥10% loss to follow-up. We categorised a trial as being at low risk of bias for a particular outcome if we identified no limitation for any risk of bias item.

Outcomes

Patients identified functional recovery (time to return to work and time to full weight bearing), pain reduction, and number of subsequent fracture or osteotomy related operations (re-operation for operatively managed fracture and osteotomy) as the most important outcomes for patients considering LIPUS for bone healing. Because many clinicians currently base their management on time to radiographic healing, a surrogate outcome important only insofar as it influences patient experience, the panel requested its inclusion in our review. We extracted all outcomes that fell into these categories as well as ultrasound device-related adverse effects.

Synthesis of results

We pooled treatment effects of LIPUS on similar outcomes across eligible trials, regardless of clinical subgroups, focusing on complete case analysis. We calculated pooled estimates and associated 95% confidence intervals (CI) using random effects models for meta-analysis with three or more studies, and fixed-effects models for meta-analysis with two studies. We examined heterogeneity associated with all pooled analyses using both the X^2 test and I^2 statistic. SAS version 9.4, R version 3.1, and Review Manager 5.3 provided software for the statistical analysis.

For time-to-event outcomes, we pooled hazard ratios. For studies that did not apply methods of survival analysis, we considered time to event reported as a continuous variable (e.g. days to return to work) at the longest follow-up time. We used the relative effect measure ratio of means (mean LIPUS/mean control) in order to account for the baseline difference in fracture healing depending on type of bone and (e.g. scaphoid, clavicle, tibia) and fracture or procedure (e.g. stress fracture or distraction osteogenesis). We pooled the natural logarithm of the ratio of means and presented the results as percentage difference (relative change). For studies that reported the proportion of patients who achieved the event at a specific time point, we calculated risk ratios.

When studies used different instruments to measure the same construct on a continuous scale, we converted all instruments to the most commonly used instrument among studies and then pooled results using the weighted mean difference.[19]

For the outcomes number of subsequent operations and device related adverse events, we calculated both risk ratios, which are preferable in case of varying baseline risks, and risk differences, which allow inclusion of studies with zero events in both groups.

In consultation with the expert and patient guideline panel, we pre-specified three subgroup hypotheses to explain heterogeneity of effects between studies: (1) LIPUS will show larger effects in high risk of bias studies, (2) LIPUS effects will differ based on clinical subgroups, and (3) LIPUS will show larger effects with greater patient compliance. In consultation with the six clinical experts on the parallel guideline panel, we classified eligible RCTs according to the following five clinical subgroups: (1) operatively managed fresh fractures, (2) non-operatively managed fresh fractures, (3) stress fractures, (4) non-union, and (5) osteotomy (including distraction osteogenesis). Because compliance was reported inconsistently, two reviewers independently categorised trials using response options of “definitely or probably high compliance” or “definitely or probably moderate compliance” using as a guide a definition of high compliance as at least 80% of patients applied LIPUS for at least 80% of the total time prescribed. We conducted univariable tests of interaction to establish if the effect size from the subgroups differed significantly from each other, and, in order to test independence of subgroup effects, performed multivariable meta-regression in which we included risk of bias (high versus low), compliance with LIPUS treatment (high versus moderate), and clinical subgroups (as above) as independent variables in a single model.

Only one outcome, days to radiographic healing, included enough studies to perform all planned subgroup analysis. As a rule of thumb we had pre-specified in our protocol at least three studies per group. We assessed the credibility of significant subgroup effects using the criteria suggested by Sun et al.[20] Based on the finding that risk of bias appeared to independently explain the high heterogeneity in the outcome days to radiographic healing, we performed subgroup analysis by risk of bias for all outcomes.

The authors and the guideline panel achieved consensus in categorising the quality of evidence for all reported outcomes as high, moderate, low, or very low using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach. In the GRADE approach, RCTs begin as high quality evidence but can be rated down due to: (1) risk of bias; (2) inconsistency; (3) indirectness; (4) imprecision; or, (5) publication bias.[21] We considered rating down for inconsistency if the magnitude and direction of effects were dissimilar, the confidence intervals had minimal overlap, the test of heterogeneity was significant, or the I^2 was high.[22] For outcomes with ten or more studies, we inspected symmetry of funnel plots and performed Egger’s statistical test for publication bias.[23]

To calculate absolute effects, we applied the effect estimate from the meta-analysis to the control arm of the TRUST trial, which enrolled patients with tibia fractures and had the largest

sample size of any eligible study that was at low risk of bias. The approach to rating certainty of individual outcomes was fully contextualised: that is, in rating quality about any individual outcome, we took into account the findings on the other outcomes.

RESULTS

Search results

We identified 3489 potentially eligible abstracts, retrieved 42 studies in full text, and found 26 eligible RCTs (fig 1).[16,24-50] Two RCTs, Handolin et al.[30,31] and Emami et al.,[27,28] provided two publications reporting on the same group of patients. There were no shared patients between the TRUST pilot [24] and the definitive trial.[16] Our registry search yielded four protocols of potentially eligible RCTs; one was discontinued due to slow recruitment (ISRCTN90844675, personal communication, outcome data not available yet), one manuscript is under peer-review (NCT00744861, personal communication: "no difference between the control group and the ultrasound group"), one is completed but unpublished (JPRN-UMIN000002005, no response from investigators), and the last is still ongoing (NCT02383160). Attempts to acquire the full text of another potentially eligible RCT,[51] reported in a recent systematic review,[11] were unsuccessful.

Study characteristics

Eligible trials enrolled patients with operatively managed fresh fractures (n=7); non-operatively managed fresh fractures (n=6); stress fractures (n=2); non-unions (n=3); and osteotomies (n=8), of which five were distraction osteogenesis (table 1). Most trials enrolled patients with tibia fractures or osteotomies (n=14). All but two trials applied LIPUS for 20 minutes every day either for a fixed period or until radiographic healing. Otherwise, one trial applied LIPUS for 15 minutes per day,[36] and another trial for 5 minutes every second day.[39] Fifteen RCTs (60%) provided their control group with an inactive device that was indistinguishable from the active LIPUS. Only three trials (12%) were explicitly free of industry funding.[26,42,48]

Risk of bias

We contacted authors to resolve areas of uncertainty and successfully clarified details in five RCTs.[30,31,35,37,40] We considered six trials to be at low risk of bias,[16,24,27,37,46,47] and the remaining 20 studies to be at high risk of bias (table 2). The main limitations were failure to report a method for allocation concealment (15 RCTs), unblinded patients (10 RCTs), caregivers or outcome assessors (10 RCTs), and high or unclear numbers of patients excluded from the analysis (13 RCTs; table 2).

Outcomes

Table 3 summarises findings of all outcomes. Interactive summary of findings tables are available online at https://www.magicapp.org/public/guideline/mL6yYj.

Functional recovery: Only the TRUST trial assessed time to return to work using a time-to-event analysis, and found no significant effect (hazard ratio 1.11 favouring control, 95% CI 0.82 to

1.50; 343 patients).[16] Three trials assessed the number of days to return to work; the pooled effect was not significant (2.7% later return with LIPUS, 95% CI 7.7% earlier to 14.3% later; I^2=0%; 392 patients) (fig 2). We found no significant interaction with risk of bias (p=0.86). Considering an alternative threshold of ≥10% loss-to follow-up for assessing risk of attrition bias, all three studies would fall into the category of high risk of bias. However, given the consist absence of effects this would not lower our confidence in the result. A fourth trial in patients with delayed union of tibia fracture provided insufficient data for inclusion in meta-analysis (table 2), but reported no significant difference in days to return to work.[50]

Only the TRUST trial assessed time to full weight bearing using a time-to event analysis, and found no significant effect (hazard ratio 0.87 in favour of LIPUS, 95% CI 0.70 to 1.08; 451 patients). Three trials assessed the number of days to full weight bearing. Overall results suggested no significant effect on full weight bearing with LIPUS but high heterogeneity (I^2=95%). The effect of the one trial at high risk of bias (40.0% earlier, 95% CI 48.4% to 30.3 earlier) differed significantly from the consistent results from the two trials at low risk of bias (4.8% later, 95% CI 4.0% earlier to 14.4% later; 483 patients; interaction p<0.001, subgroup effect not effected by alternative threshold for missing data) (fig 3).

Appendix 2 presents results of other functional outcomes including return to leisure activities, return to household activities, return to pre-injury level of function, and physical function measured with a multidimensional questionnaire. None of these were significantly affected by use of LIPUS, nor did they show substantial inconsistency.

Pain reduction: Four trials assessed pain, two using a 100mm visual analogue scale[37,49] and two using the subdomain "bodily pain" of the SF-36 instrument.[16,24] After transforming all results to a 100mm visual analogue scale, findings at 3 to 6 weeks follow-up showed no significant effect of LIPUS on pain reduction but high heterogeneity (I^2=97%). The effect of the one trial at high risk of bias (28.12 mm lower, 95% CI, 37.05 to 19.19 lower) differed significantly from the consistent results from the three trials at low risk of bias (0.93 mm lower, 95% CI 2.51 lower to 0.64 higher; 626 patients; I^2=0%; interaction p<0.001; fig 4). The subgroup effect was no longer significant when we used a threshold of ≥10% missing data to designate a trial at high risk of attrition bias (p=0.35, fig 4 in Appendix 4). Two other small studies assessed pain intensity at 5 months but could not be included in the meta-analysis. One reported pain outcomes only narratively (no effect),[41] another used a modified instrument with unclear scale and variance (no effect).[49]

Other outcomes for pain included pain intensity assessed at multiple time-points and number of painful days (appendix 3). None showed a significant effect of LIPUS, nor substantial inconsistency.

Number of subsequent operations: Ten trials reported the number of subsequent operations including three trials reporting zero events in both arms. Neither the pooled risk ratio (0.8 in favour of LIPUS, 95% CI 0.55 to 1.16; I^2=0%; 7 trials, 693 patients; fig 5) nor the pooled risk difference (3% reduction with LIPUS, 95% CI 7% reduction to 2% increase; I^2=0%; 10 trials, 740

patients) showed a significant effect. There was no significant interaction with risk of bias on either scale (risk ratio: p=0.75; risk difference: p=0.64. The results did not depend on the threshold for missing data).

Time to radiographic healing: Two trials used time-to-event analysis methods to assess time to radiographic healing,[16,24] and showed no significant effect of LIPUS (hazard ratio 1.06 in favour of control, 95% CI 0.86 to 1.32; I^2=0%; 532 patients). Fifteen trials reported the number of days to radiographic healing. Overall results suggested accelerated radiographic healing with LIPUS (26% earlier, 95% CI 33.6% to 17.8% earlier; I^2=84.7%). The effect differed significantly between the 12 trials at high risk of bias (31.8% earlier; 95%CI 38.6% to 24.3% days earlier; I^2=77.8%; 446 patients) and the three trials at low risk of bias (1.7% earlier, 95% CI 11.2% earlier to 8.8% later, I^2=9.8%; 483 patients; interaction p<0.001; fig 6). This subgroup effect fulfilled 8 of 9 credibility criteria relevant to risk of bias as an explanation of heterogeneity (table 4). In addition, the subgroup effect was robust to our sensitivity analysis using a more conservative threshold for defining risk of attrition bias (interaction p=0.004, fig 5 in appendix 4). The effect of LIPUS on days to radiographic healing did not differ significantly across clinical subgroups (p=0.13, fig 1 in appendix 4) or between high and moderate compliance with treatment (p=0.79, fig 2 in appendix 4). In our multivariable meta-regression, which included risk of bias, clinical subgroups, and compliance with treatment, the only significant effect modifier was the risk of bias (p=0.005).

Another RCT in patients with delayed union of tibia fracture reported only the proportion of healed fractures at 16 weeks and did not find a significant difference (65% in the LIPUS and 46% in the control arm, p=0.07; high risk of bias towards LIPUS due to serious imbalance in age of fracture at baseline).[44]

The funnel plot based on time to radiographic healing was not clearly asymmetrical and Egger's test for publication bias was not significant (p=0.25, fig 3 in appendix 4).

Device related adverse effects: Seven studies reported explicitly the absence of any device-related adverse effects; two other studies reported mild transient skin irritations in 6 of patients. The pooled risk ratio based on these two studies (2.65 in favour of control, 95% CI 0.32 to 22.21; 129 patients) was not significant, nor was the pooled risk difference based on all nine trials (0%, 95% CI 1% reduction to 1% increase; I^2=0%; 839 patients; fig 7). We found no significant interaction with risk of bias on the risk difference scale (p=0.75).

DISCUSSION

Main findings

Our systematic review demonstrated moderate quality evidence that LIPUS applied to patients with fractures or osteotomies has no effect on time to return to work or the number of subsequent operations (table 3). Overall results suggested a possible reduction of days to full weight bearing, pain, and days to radiographic healing, but with large variability between

studies strongly associated with risk of bias as an effect modifier: only trials with high risk of bias demonstrated benefit. Based on RCTs at low risk of bias, we found high quality evidence that LIPUS has no effect on pain reduction, days to full weight bearing, or device-related adverse effects, and moderate quality evidence that LIPUS has no effect on days to radiographic healing (table 3).

Comparison with other systematic reviews

Our results are consistent with other systematic reviews in concluding that most RCTs addressing LIPUS therapy are poorly reported, lack patient important outcomes, and are at high risk of bias.[5-14] Our systematic review, however, differs from previous systematic reviews in several important aspects. First, we include the recently published TRUST trial,[16] by far the largest trial addressing LIPUS therapy for bone healing, which reported a number of patient-important outcomes. Second, our choice of outcomes and interpretation of findings was informed by a guideline panel including patients with personal experience of fractures in the context of BMJ Rapid Recommendations. Patients considered functional recovery, pain reduction and operations as critical outcomes, while expressing little interest in the commonly reported surrogate outcome of radiographic healing. Third, we used optimal statistical approaches, and in particular the effect measure *ratio of means* (rather than difference of means) to combine days to radiographic healing, return to work, or full weight bearing across studies. This relative effect measure is most appropriate in the context of LIPUS where the average time to recovery differs substantially between clinical subgroups. For instance, a lower grade stress fracture is likely to heal much faster than a complicated tibia fracture. It is not surprising, therefore, that previous meta-analyses found high heterogeneity when they used absolute mean differences to pool across studies.[8,11,12]

Finally, we used the GRADE approach to assess the quality of evidence, taking into account the results of subgroup analysis based on risk of bias: when effects differed significantly between high and low quality trials, we based our conclusions on trials at low risk of bias. Our approach of limiting conclusions to low risk of bias trials depends on our judgement of risk of bias; however, our ratings of risk of bias were consistent with those of a previous Cochrane systematic review.[5] Further, most trials judged to be at high risk of bias had limitations in more than one domain, and some had additional sources of bias including baseline imbalance or unclear clustering when patients had more than one fracture or surgery. Applying our risk of bias judgments as an effect modifier met 8 of 9 relevant criteria for a credible subgroup analysis (table 4). A post-hoc sensitivity analysis exploring a more conservative threshold for attrition bias (≥10% loss to follow-up) yielded, for all outcomes, results essentially consistent with the primary analyses

Limitations

The primary limitation of our review is the failure of most trials to measure or report patient-important outcomes. Of the 26 eligible trials, 11 reported, in sufficient detail for inclusion in meta-analysis, outcomes that patient consider critical for decision making.[16,24,25,27,29-

[31,35,37,39,46,47] Of these, the only four trials that contributed substantial data were either conducted in patients with operatively managed fresh tibia fracture[16,24,27] or conservatively managed clavicle fracture.[37] One could question the extent to which our results apply to patients not included at all (such as children) or underrepresented (stress fractures, non-union, and osteotomies) in the eligible trials. Qualitative subgroup effects (e.g. no benefit in one subgroup and important benefit in another) are, however, unusual. In the absence of evidence to the contrary, it is therefore reasonable to apply our results to these populations. Our subgroup analysis and meta-regression for radiographic healing found no effect modification based on clinical subgroups. Certainly, the burden of proof regarding the effect of LIPUS in children and underrepresented populations rests with those who might postulate a benefit.

LIPUS compared with electrical stimulation

Our findings are similar to a 2016 systematic review of 15 small trials that explored electrical stimulation vs. sham therapy for fracture healing; only 4 of which were at low risk of bias.[52] This review found moderate quality evidence for a 35% reduction (95% CI 19% to 47%; I^2=46%) in the rate of radiographic nonunion. The authors found no evidence of a subgroup effect based on clinical presentation (i.e. fresh fractures, delayed union or nonunion, spinal fusion, or surgical osteotomy; interaction p = 0.41) – they did not explore whether risk of bias explained heterogeneity, but all 4 trials at low risk of bias showed no significant effect on radiographic union.[53-56] This review found a small reduction in pain (mean difference of −7.7 mm on a 100mm visual analogue scale for pain, 95% CI −13.92 to −1.43), and low quality evidence for no difference in functional outcome (mean difference of −0.88 points on the 100 point Short Form 36 Physical Component Summary score, 95% CI −6.63 to 4.87).

Conclusions

In conclusion, moderate to high quality evidence demonstrates that LIPUS fails to accelerate return to work, return to full weight bearing, pain, or the need for subsequent operation. If one gives highest credibility to combined effects from all available RCTs, low quality evidence would suggest a large reduction in time to radiographic healing. If, however, one gives higher credence to low risk of bias trials, moderate to high quality evidence suggests that LIPUS not only has no effect on patient-important outcomes, but also fails to accelerate radiographic healing. The evidence applies directly to patients with fresh fractures and indirectly to children and other underrepresented populations, particularly those with non-union, for which no trustworthy direct evidence exists.

Acknowledgements

We thank members of the Rapid Recommendations panel for critical feedback on outcome and subgroup selection, GRADE judgments, and manuscript feedback, including Rudolf Poolman (chair and orthopaedic surgeon), Ian Harris (orthopaedic and trauma surgeon), Inger Schipper (trauma surgeon), Maureen Smith (patient representative), Alexandra Albin (patient representative), Sally Nador (patient representative), William Sasges (patient representative), Ton Kuijpers (methodologist), Loes van Beers (physiotherapist), and Michael Verhofstad (trauma surgeon).

Competing interest
All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work. We acknowledge that JWB, DHA, and GHG were co-authors of the TRUST trial, which was supported in part by an industry grant from Smith & Nephew, a manufacturer of LIPUS devices. No other relationships or activities that could appear to have influenced the submitted work.

Funding
This study was unfunded.

Contributors
JWB, RAS, GHG and POV conceived the study idea. SS and JWB coordinated the systematic review. SS wrote the first draft of the manuscript. RC designed the search strategy. LL, AK, RC, and SS screened abstracts and full texts. LL, AK, RAS, TA, and SS acquired the data and judged risk of bias in the studies. SS, JWB and DHA performed the data analysis. DHA and GHG provided statistical advice. SS, RAS, POV, JWB and GHG interpreted the data analysis. All authors critically revised the manuscript. SS had full access to all of the data in the study, and takes responsibility for the integrity of the data and the accuracy of the data analysis. SS is the guarantor.

Ethical approval
Not required.

Data sharing
All data informing the study is freely available in the appendices.

Data access
All authors had full access to all of the data (including statistical reports and tables) in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.

Patient involvement
Four patient representatives were full members of the guideline, and contributed to the selection and prioritisation of outcomes, values and preferences assessments, and critical feedback to the protocol for the systematic review and the BMJ Rapid Recommendations manuscript.

Transparency declaration
SS is guarantor and affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

REFERENCES

1.National Institute for Helacare Excellence. **Low-intensity pulsed ultrasound to promote fracture healing.** 2010. URL: https://www.nice.org.uk/guidance/ipg374/history.
2.Food and Drug Administration. **Approval order for Exogen device.** 2000, URL: http://www.fda.gov/ohrms/dockets/dailys/00/mar00/031300/aav0001.pdf.
3.Busse JW, Morton E, Lacchetti C, Guyatt GH, Bhandari M. **Current management of tibial shaft fractures: a survey of 450 Canadian orthopedic trauma surgeons.** Acta Orthop. 2008;79(5):689-94.
4.Wachovia Capital Markets. **Equity research: bone growth stimulation 2008 outlook.** 2007.
5.Griffin XL, Parsons N, Costa ML, Metcalfe D. **Ultrasound and shockwave therapy for acute fractures in adults.** Cochrane Database Syst Rev. 2014(6):CD008579.
6.Ebrahim S, Mollon B, Bance S, Busse JW, Bhandari M. **Low-intensity pulsed ultrasonography versus electrical stimulation for fracture healing: a systematic review and network meta-analysis.** Can J Surg. 2014;57(3):E105-18.
7.Bashardoust Tajali S, Houghton P, MacDermid JC, Grewal R. **Effects of low-intensity pulsed ultrasound therapy on fracture healing: a systematic review and meta-analysis.** Am J Phys Med Rehabil. 2012;91(4):349-67.
8.Hannemann PF, Mommers EH, Schots JP, Brink PR, Poeze M. **The effects of low-intensity pulsed ultrasound and pulsed electromagnetic fields bone growth stimulation in acute fractures: a systematic review and meta-analysis of randomized controlled trials.** Arch Orthop Trauma Surg. 2014;134(8):1093-106.
9.Martinez de Albornoz P, Khanna A, Longo UG, Forriol F, Maffulli N. **The evidence of low-intensity pulsed ultrasound for in vitro, animal and human fracture healing.** Br Med Bull. 2011;100:39-57.
10.Raza H, Saltaji H, Kaur H, Flores-Mir C, El-Bialy T. **Effect of Low-Intensity Pulsed Ultrasound on Distraction Osteogenesis Treatment Time: A Meta-analysis of Randomized Clinical Trials.** J Ultrasound Med. 2016;35(2):349-58.
11.Rutten S, van den Bekerom MP, Sierevelt IN, Nolte PA. **Enhancement of Bone-Healing by Low-Intensity Pulsed Ultrasound: A Systematic Review.** JBJS Rev. 2016;4(3).
12.Snyder BM, Conley J, Koval KJ. **Does low-intensity pulsed ultrasound reduce time to fracture healing? A meta-analysis.** Am J Orthop (Belle Mead NJ). 2012;41(2):E12-9.
13.Watanabe Y, Matsushita T, Bhandari M, Zdero R, Schemitsch EH. **Ultrasound for fracture healing: current evidence.** J Orthop Trauma. 2010;24 Suppl 1:S56-61.
14.Busse JW, Kaur J, Mollon B, Bhandari M, Tornetta P, Schunemann HJ, et al. **Low intensity pulsed ultrasonography for fractures: systematic review of randomised controlled trials.** Bmj. 2009;338:b351.
15.Siemieniuk RA, Agoritsas T, Macdonald H, Guyatt GH, Brandt L, Vandvik PO. **Introduction to BMJ Rapid Recommendations.** BMJ. 2016;354:i5191.
16.TRUST Investigators writing group, Busse JW, Bhandari M, Einhorn TA, Schemitsch E, Heckman JD, et al. **Re-evaluation of low intensity pulsed ultrasound in treatment of tibial fractures (TRUST): randomized clinical trial.** Bmj. 2016;355:i5351.
17.Poolman RW, Agoritsas T, Siemieniuk RAC, Harris IA, Schipper IB, Mollon B, et al. **Low intensity pulsed ultrasound (LIPUS) for bone healing: a clinical practice guideline.** Submitted to The BMJ 2016.

18.Akl EA, Sun X, Busse JW, Johnston BC, Briel M, Mulla S, et al. **Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid.** J Clin Epidemiol. 2012;65(3):262-7.
19.Thorlund K, Walter SD, Johnston BC, Furukawa TA, Guyatt GH. **Pooling health-related quality of life outcomes in meta-analysis-a tutorial and review of methods for enhancing interpretability.** Res Synth Methods. 2011;2(3):188-203.
20.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.
21.Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. **GRADE: an emerging consensus on rating quality of evidence and strength of recommendations.** BMJ. 2008;336(7650):924-6.
22.Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. **GRADE guidelines: 7. Rating the quality of evidence--inconsistency.** J Clin Epidemiol. 2011;64(12):1294-302.
23.Egger M, Davey Smith G, Schneider M, Minder C. **Bias in meta-analysis detected by a simple, graphical test.** BMJ. 1997;315(7109):629-34.
24.Busse JW, Bhandari M, Einhorn TA, Heckman JD, Leung KS, Schemitsch E, et al. **Trial to re-evaluate ultrasound in the treatment of tibial fractures (TRUST): a multicenter randomized pilot study.** Trials. 2014;15:206.
25.Dudda M, Hauser J, Muhr G, Esenwein SA. **Low-intensity pulsed ultrasound as a useful adjuvant during distraction osteogenesis: a prospective, randomized controlled trial.** Journal of Trauma-Injury Infection & Critical Care. 2011;71(5):1376-80.
26.El-Mowafi H, Mohsen M. **The effect of low-intensity pulsed ultrasound on callus maturation in tibial distraction osteogenesis.** Int Orthop. 2005;29(2):121-4.
27.Emami A, Petren-Mallmin M, Larsson S. **No effect of low-intensity ultrasound on healing time of intramedullary fixed tibial fractures.** J Orthop Trauma. 1999;13(4):252-7.
28.Emami A, Larsson A, Petren-Mallmin M, Larsson S. **Serum bone markers after intramedullary fixed tibial fractures.** Clin Orthop Relat Res. 1999(368):220-9.
29.Gan TY, Kuah DE, Graham KS, Markson G. **Low-intensity pulsed ultrasound in lower limb bone stress injuries: a randomized controlled trial.** Clin J Sport Med. 2014;24(6):457-60.
30.Handolin L, Kiljunen V, Arnala I, Kiuru MJ, Pajarinen J, Partio EK, et al. **Effect of ultrasound therapy on bone healing of lateral malleolar fractures of the ankle joint fixed with bioabsorbable screws.** J Orthop Sci. 2005;10(4):391-5.
31.Handolin L, Kiljunen V, Arnala I, Pajarinen J, Partio EK, Rokkanen P. **The effect of low intensity ultrasound and bioabsorbable self-reinforced poly-L-lactide screw fixation on bone in lateral malleolar fractures.** Arch Orthop Trauma Surg. 2005;125(5):317-21.
32.Handolin L, Kiljunen V, Arnala I, Kiuru MJ, Pajarinen J, Partio EK, et al. **No long-term effects of ultrasound therapy on bioabsorbable screw-fixed lateral malleolar fracture.** Scand J Surg. 2005;94(3):239-42.
33.Heckman JD, Ryaby JP, McCabe J, Frey JJ, Kilcoyne RF. **Acceleration of tibial fracture-healing by non-invasive, low-intensity pulsed ultrasound.** J Bone Joint Surg Am. 1994;76(1):26-34.
34.Kristiansen TK, Ryaby JP, McCabe J, Frey JJ, Roe LR. **Accelerated healing of distal radial fractures with the use of specific, low-intensity ultrasound. A multicenter, prospective, randomized, double-blind, placebo-controlled study.** J Bone Joint Surg Am. 1997;79(7):961-73.

35.Leung KS, Lee WS, Tsui HF, Liu PP, Cheung WH. **Complex tibial fracture outcomes following treatment with low-intensity pulsed ultrasound.** Ultrasound Med Biol. 2004;30(3):389-95.
36.Liu Y, Wei X, Kuang Y, Zheng Y, Gu X, Zhan H, et al. **Ultrasound treatment for accelerating fracture healing of the distal radius. A control study.** Acta Cir Bras. 2014;29(11):765-70.
37.Lubbert PH, van der Rijt RH, Hoorntje LE, van der Werken C. **Low-intensity pulsed ultrasound (LIPUS) in fresh clavicle fractures: a multi-centre double blind randomised controlled trial.** Injury. 2008;39(12):1444-52.
38.Mayr E, Rudzki MM, Rudzki M, Borchardt B, Hausser H, Ruter A. **Does pulsed low-intensity ultrasound accelerate healing of scaphoid fractures?. [German].** Handchirurgie Mikrochirurgie Plastische Chirurgie. 2000;32(2):115-22.
39.Patel K, Kumar S, Kathiriya N, Madan S, Shah A, Venkataraghavan K, et al. **An Evaluation of the Effect of Therapeutic Ultrasound on Healing of Mandibular Fracture.** Craniomaxillofac Trauma Reconstr. 2015;8(4):299-306.
40.Ricardo M. **The effect of ultrasound on the healing of muscle-pediculated bone graft in scaphoid non-union.** Int Orthop. 2006;30(2):123-7.
41.Rutten S, Klein-Nulend J, Guit GL, Albers GHR, Korstjens CM, M. WPIJ, et al. Use of low-intensity pulsed ultrasound stimulation of delayed unions of the osteotomized fibula: a prospective randomized double-blind trial (Thesis). Low-intensity pulsed ultrasound treatment in delayed bone healing [thesis]. Amsterdam: Vrije Universiteit Amsterdam; 2012.
42.Rue JP, Armstrong DW, 3rd, Frassica FJ, Deafenbaugh M, Wilckens JH. **The effect of pulsed ultrasound in the treatment of tibial stress fractures.** Orthopedics. 2004;27(11):1192-5.
43.Salem KH, Schmelz A. **Low-intensity pulsed ultrasound shortens the treatment time in tibial distraction osteogenesis.** International Orthopaedics. 2014;38(7):1477-82.
44.Schofer MD, Block JE, Aigner J, Schmelz A. **Improved healing response in delayed unions of the tibia with low-intensity pulsed ultrasound: results of a randomized sham-controlled trial.** BMC Musculoskelet Disord. 2010;11:229.
45.Kamath JB, Jayasheelan N, Reddy B, Muhammed S, Savur A. **The effect of low-intensity pulsed ultrasound therapy on fracture healing.** Muller Journal of Medical Sciences and Research. 2015;6(1):49-53.
46.Schortinghuis J, Bronckers AL, Stegenga B, Raghoebar GM, de Bont LG. **Ultrasound to stimulate early bone formation in a distraction gap: a double blind randomised clinical pilot trial in the edentulous mandible.** Arch Oral Biol. 2005;50(4):411-20.
47.Schortinghuis J, Bronckers AL, Gravendeel J, Stegenga B, Raghoebar GM. **The effect of ultrasound on osteogenesis in the vertically distracted edentulous mandible: a double-blind trial.** Int J Oral Maxillofac Surg. 2008;37(11):1014-21.
48.Tsumaki N, Kakiuchi M, Sasaki J, Ochi T, Yoshikawa H. **Low-intensity pulsed ultrasound accelerates maturation of callus in patients treated with opening-wedge high tibial osteotomy by hemicallotasis.** J Bone Joint Surg Am. 2004;86-a(11):2399-405.
49.Urita A, Iwasaki N, Kondo M, Nishio Y, Kamishima T, Minami A. **Effect of low-intensity pulsed ultrasound on bone healing at osteotomy sites after forearm bone shortening.** Journal of Hand Surgery – American Volume. 2013;38(3):498-503.
50.Zacherl M, Gruber G, Radl R, Rehak PH, Windhager R. **No midterm benefit from low intensity pulsed ultrasound after chevron osteotomy for hallux valgus.** Ultrasound Med Biol. 2009;35(8):1290-7.

51.Nolte PA, Maas M, Roolker L, Marti RK, Albers GHR. Effect of low-intensity ultrasound on bone healing in osteotomies of the lower extremity: a randomised trial. In: Nolte PA, editor. Nonunions – surgery and low-intensity ultrasound treatment [thesis]. Amsterdam: Universiteit van Amsterdam; 2002. p. 96-106.
52.Aleem IS, Aleem I, Evaniew N, Busse JW, Yaszemski M, Agarwal A, et al. **Efficacy of Electrical Stimulators for Bone Healing: A Meta-Analysis of Randomized Sham-Controlled Trials.** Sci Rep. 2016;6:31724.
53.Hannemann PF, Gottgens KW, van Wely BJ, Kolkman KA, Werre AJ, Poeze M, et al. **The clinical and radiological outcome of pulsed electromagnetic field treatment for acute scaphoid fractures: a randomised double-blind placebo-controlled multicentre trial.** J Bone Joint Surg Br. 2012;94(10):1403-8.
54.Hannemann PF, van Wezenbeek MR, Kolkman KA, Twiss EL, Berghmans CH, Dirven PA, et al. **CT scan-evaluated outcome of pulsed electromagnetic fields in the treatment of acute scaphoid fractures: a randomised, multicentre, double-blind, placebo-controlled trial.** Bone Joint J. 2014;96-B(8):1070-6.
55.Mammi GI, Rocchi R, Cadossi R, Massari L, Traina GC. **The electrical stimulation of tibial osteotomies. Double-blind study.** Clin Orthop Relat Res. 1993(288):246-53.
56.Martinez-Rondanelli A, Martinez JP, Moncada ME, Manzi E, Pinedo CR, Cadavid H. **Electromagnetic stimulation as coadjuvant in the healing of diaphyseal femoral fractures: a randomized controlled trial.** Colomb Med (Cali). 2014;45(2):67-71.

Table 1: Study characteristics

Author Year	Bone	Type of fracture / surgery	% open fracture	Manage-ment	% women	Mean age	N randomised		Sham device	Dose and duration of LIPUS therapy	Max follow-up	Explicit free of industry funding
							LIPUS	No ultrasound				
Busse 2014(24)	Tibia	Fresh fracture	27%	Operative	24%	40	23	28	Yes	20 min/day to healing*	1 year	No
Busse 2016(16)	Tibia	Fresh fracture	23%	Operative	31%	40	250	251	Yes	20 min/day to healing*	1 year	No
Dudda 2011(25)	Tibia	Distraction osteogenesis	NA	Operative	11%	39	16	20	No	20 min/day to healing*	35 weeks	No
El-Mowafi 2005(26)	Tibia	Distraction osteogenesis	NA	Operative	0%	35	10	10	No	20 min/day to healing*	12 months	Yes
Emami 1999(27, 28)	Tibia	Fresh fracture	13%	Operative	25%	37	15	17	Yes	20 min/day to healing*	20 weeks	No
Gan 2014(29)	Tibia, fibula, metatarsal	Stress fracture	0%	Non-operative	83%	30	15	15	Yes	20 min/day for 28 days	12 weeks	No
Handolin 2005a(30, 31)	Lateral malleolus	Fresh fracture	0%	Operative	47%	42	11	11	Yes	20 min/day for 42 days	12 weeks	No
Handolin 2005b (32)	Lateral malleolus	Fresh fracture	0%	Operative	56%	40	15	15	Yes	20 min/day for 42 days	18 months	No
Heckman 1994(33)	Tibia	Fresh fracture	4%	Non-operative	19%	33	48	49	Yes	20 min/day to healing*	140 days	No
Kamath 2015(45)	Tibia and femur	Fresh fracture	0%	Operative	NR	36	33	27	No	20 min/day for 1 month	16 weeks	No
Kristiansen 1997(34)	Distal radius	Fresh fracture	0%	Non-operative	84%	56	40	45	Yes	20 min/day for 70 days	140 days	No
Leung 2004(35)	Tibia	Fresh fracture	47%	Operative	11%	35	16	14	Yes	20 min/day for 4 months	5 months	No
Liu 2014(36)	Distal radius	Fresh fracture	NR	Non-operative	36%	67	41	40	No	15 min/day for ≥12 weeks	At least 12 weeks	No
Lubbert 2008(37)	Clavicle	Fresh fracture	0%	Non-operative	16%	38	61	59	Yes	20 min/day for 28 days	8 weeks	No
Mayr 2000(38)	Scaphoid	Fresh fracture	0%	Non-operative	17%	37	15	15	No	20 min/day to healing*	120 days	No
Patel 2014(39)	Mandible	Fresh fracture	NR	Non-operative	25%	15-35	14	14	No	5 min q.a.d. for 24 days	5 weeks	No

Table 1: Study characteristics (continued)

Ricardo 2006(40)	Scaphoid	Non-union	NA	Operative	0%	27	10	11	Yes	20 min/day to healing*	4 years	No
Rue 2004(42)	Tibia	Stress fracture	0%	Non-operative	50%	19	Probably 20	Probably 20	Yes	20 min/day to healing*	NR	Yes
Rutten 2012(41)	Tibia	Non-union	0%	Operative	70%	41-63	10	10	Yes	20 min/day for 5 months	5 years	No
Salem 2014(43)	Tibia	Distraction osteogenesis	NA	Operative	14%	30	12	9	No	20 min/day to healing*	NR	No
Schofer 2010(44)	Tibia	Non-union	NA	Operative	24%	44	51	50	Yes	20 min/day for 16 weeks	16 weeks	No
Schortinghuis 2005(46)	Mandible	Distraction osteogenesis	NA	Operative	75%	65	4	4	Yes	20 min/day for 4 weeks	30 months	No
Schortinghuis 2008(47)	Mandible	Distraction osteogenesis	NA	Operative	NR	56	5	4	Yes	20 min/day for 6 weeks	44 months	No
Tsumaki 2004(48)	Tibia	Distraction osteogenesis	NA	Operative	81%	68	21 knees	21 knees	No	20 min/day to healing*	NR	Yes
Urita 2013(49)	Ulna and radius	Osteotomy (shortening)	NA	Operative	63%	48	14	13	No	20 min/day to healing* or 12 weeks	24 weeks	No
Zacherl 2009(50)	Hallux valgus	Osteotomy (deformity correction)	NA	Operative	85%	53	26 toes	26 toes	Yes	20 min/day for 42 days	1 year	No

*Until radiographic healing. q.a.d . = every other day

Table 2: Risk of bias

Author Year	Sequence generation adequate	Concealment of treatment allocation	Patients blinded	Caregivers blinded	Outcome assessors blinded	Outcomes reported as planned (link to protocol)	No other bias detected	Loss to follow-up (%) for outcome radiographic healing unless specified otherwise
Busse 2014(24)	Yes	Yes	Yes	Yes	Yes	Yes[a]	Yes	2%
Busse 2016(16)	Yes	Yes	Yes	Yes	Yes	Yes[a]	Yes	19% for radiographic healing, 11% for return to work, 9% for weight bearing
Dudda 2011(25)	Yes	No	No	No	No	Unclear[b]	Yes	Unclear, assumed to be 0
El-Mowafi 2005(26)	Yes	No	No	No	No	Unclear[b]	Yes	5%
Emami 1999(27, 28)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	Yes	3%
Gan 2014(29)	Yes	No	Yes	Yes	Yes	Unclear[b]	Yes	23% (pain)
Handolin 2005a(30, 31)	Yes	No	Yes	Yes	Yes	Unclear[b]	Yes	5%
Handolin 2005b (32)	Yes	No	Yes	Yes	Yes	Unclear[b]	Yes	No eligible outcome reported
Heckman 1994(33)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	Yes	31%
Kamath 2015(45)	Yes	No	No	No	Yes	Unclear[b]	Yes	No eligible outcome reported
Kristiansen 1997(34)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	Yes	28%
Leung 2004(35)	No[c]	No[c,d]	No[d]	No[d]	No[d]	Unclear[b]	No[e]	Unclear, assumed to be 0
Liu 2014(36)	Yes	No	No	No	yes	Unclear[b]	No[f]	Unclear, assumed to be 0
Lubbert 2008(37)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	Yes	16%
Mayr 2000(38)	Yes	No	No	No	Yes	Unclear[b]	Yes	0
Patel 2014(39)	Yes	No	No	No	No	Unclear[b]	Yes	Unclear, assumed to be 0

Table 2: Risk of bias (continued)

Ricardo 2006(40)	Yes	No	Yes	Yes	Yes	Unclear[b]	Yes	Unclear, assumed to be 0
Rue 2004(42)	Yes	No	Yes	Yes	Yes	Unclear[b]	Yes	Unclear, probably 35%
Rutten 2012(41)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	Yes	45%
Salem 2014(43)	Yes	No	No	No	No	Unclear[b]	Yes	Unclear, assumed to be 0
Schofer 2010(44)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	No[g]	Unclear, assumed to be 0
Schortinghuis 2005(46)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	Yes	0 for subsequent operation
Schortinghuis 2008(47)	Yes	Yes	Yes	Yes	Yes	Unclear[b]	Yes	0 for subsequent operation
Tsumaki 2004(48)	Yes	Yes	No	No	No	Unclear[b]	No[h]	Unclear, assumed to be 0
Urita 2013(49)	No[i]	No	No	No	Yes	Unclear[b]	Yes	Unclear, assumed to be 0
Zacherl 2009(50)	Yes	No	yes	yes	yes	Unclear[b]	No[k]	Not included in meta-analysis due to insufficient reporting[k]

a Protocol: NCT00667849
b No protocol published and trial not registered
c Quasi-randomised based on sequence of admission
d Inactive device was distinguishable from active device
e Unadjusted clustering, 30 fractures of 28 patients were randomized
f Implausibly narrow confidence intervals
g Prognostic imbalance: non-union fractures in LIPUS arm were considerably older
h Bilateral surgery – one tibia was randomised to LIPUS and one to no treatment. We assumed a correlation of 0.5 in our analysis of days to radiographic healing
i Used an odd-even system for treatment allocation
k Randomised 44 patients but analysed 52 toes, clustering unclear, standard deviations not reported

Table 3: GRADE Summary of Findings table

Outcome	Study results and measurements	Absolute effect estimates		Quality of evidence	Narrative Summary
		No ultrasound	LIPUS		
Days to return to work	% Difference: 2.7% (95% CI, -7.7% to 14.3%) in days, lower better, based on data from 392 patients in 3 studies	**200 days** (Mean)	**205 days** (Mean)	**Moderate** Due to serious imprecision	LIPUS probably has little or no impact on time to return to work
		Difference: 5 days later (95% CI, 15 earlier to 20 later)			
Days to full weight bearing	% Difference: 4.8% (95% CI, -4.0% to 14.4%) in days, lower better, based on data from 483 patients in 2 trials at low risk of bias	**70 days** (Mean)	**73 days** (Mean)	**High**	LIPUS has no impact on time to full weight bearing
		Difference: 3 days earlier (95% CI, 3 earlier to 10 later)			
Pain reduction Follow up 4 to 6 weeks	Mean difference: -0.93 (95% CI -2.51 to 0.64) 0 to 100 visual analogue scale, lower better, minimal important difference: 10-15, based on data from 626 patients in 3 trials at low risk of bias	**40** (Mean)	**39** (Mean)	**High**	LIPUS has no impact on pain reduction
		Difference: 1 lower (95% CI 3, lower to 1 higher)			
Subsequent operations Follow up 8 weeks to 44 months	Risk ratio: 0.80 (95% CI 0.55 to 1.16) Based on data from 740 patients in 7 studies	**160 per 1000**	**128 per 1000**	**Moderate** Due to serious imprecision	LIPUS probably has little or no impact on subsequent operation
		Difference: 32 fewer (95% CI, 72 fewer to 26 more)			
Days to radiographic healing	% Difference: -1.7% (95% CI, -11.2% to 8.8%) in days, lower better, based on data from 483 patients in 3 trials at low risk of bias	**150 days (Mean)**	**147 days (Mean)**	**Moderate** Due to serious imprecision	LIPUS probably has little or no impact on time to radiographic healing
		Difference: 3 days earlier (95% CI, 17 earlier to 13 later)			
Device-related adverse effects Follow up 5 to 52 weeks	Risk difference: 0% (CI 95% -1% to 1%) Based on data from 839 patients in 9 studies	**0 per 1000**	**0 per 1000**	**High**	LIPUS has no impact on device-related adverse effects
		Difference: 0 fewer (95% CI, 10 fewer to 10 more)			

Table 4: Credibility of subgroup effects for risk of bias for the outcome days to radiographic healing

Criteria(20)	Rating (yes means higher credibility)
Is the subgroup variable a characteristic measured at baseline or after randomization?	Not applicable for risk of bias
Is the effect suggested by comparisons within rather than between studies?	No, between studies
Was the subgroup effect specified a priori?	Yes, specified in our protocol
Was the direction of the subgroup effect specified a priori?	Yes, we expected a larger effect for studies at high risk of bias
Is there indirect evidence that supports the hypothesized interaction (biological rationale)?	Not applicable for risk of bias
Was the subgroup effect one of a small number of hypothesized effects tested?	Yes, one of three
Does the interaction test suggest a low likelihood that chance explains the apparent subgroup effect?	Yes, significant in univariable subgroup analysis ($p<0.001$)
Is the significant subgroup effect independent?	Yes, significant in multivariable meta-regression ($p<0.01$)
Is the size of the subgroup effect large?	Yes, 31.8% acceleration in high risk of bias trials versus 1.7% acceleration in low risk of bias trials
Is the interaction consistent across closely related outcomes within the study?	Yes, risk of bias explained heterogeneity in outcomes weight bearing and pain
Is the interaction consistent across studies?	Yes, high risk of bias studies consistently showed large effects, low risk of bias studies small effects

Figure 1: Flow diagram of studies included in review of low intensity pulsed ultrasound compared with control (sham device or no device) for patients with fracture or osteotomy

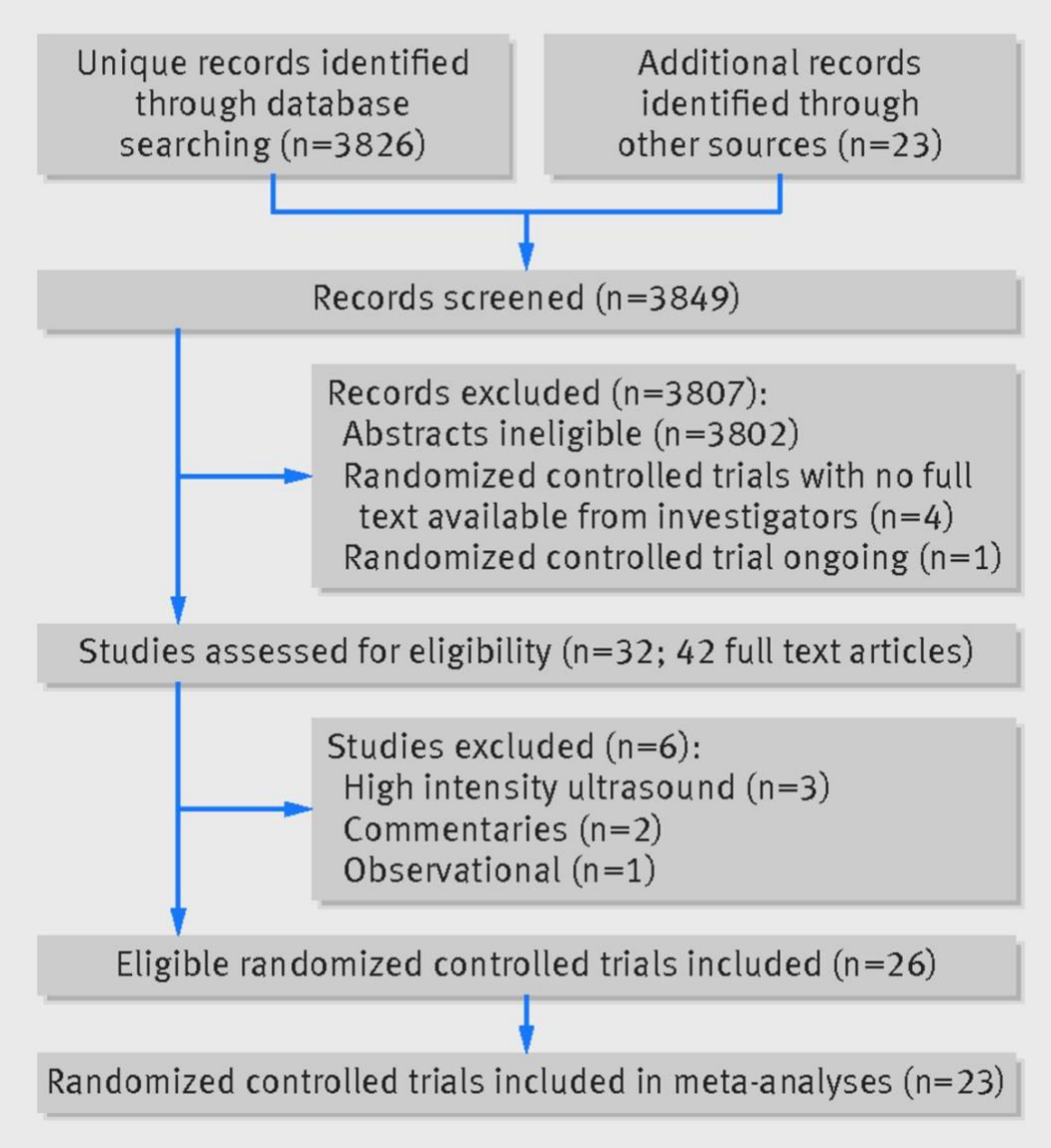


Figure 2: Forest plot for percent difference of days to return to work for low intensity pulsed ultrasound (LIPUS) compared with control (sham device or no device)

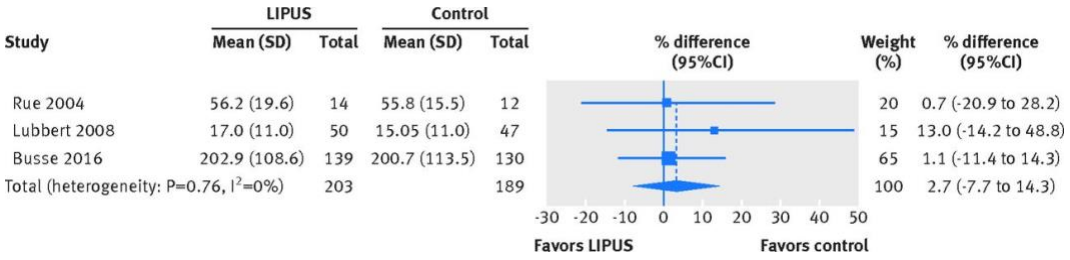


Figure 3: Forest plot for percent difference of days to full weight bearing for low intensity pulsed ultrasound (LIPUS) compared with control (sham device or no device), by risk of bias. Interaction p<0.001

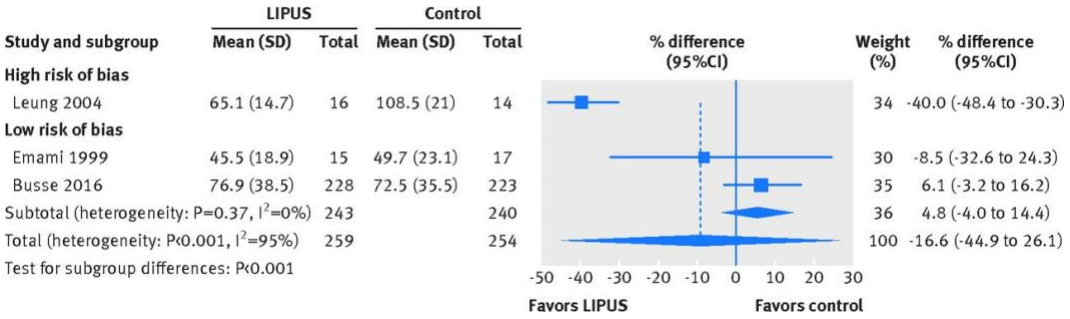


Figure 4: Forest plot for mean difference of pain reduction, all instruments transformed to 0-100 visual analogue scale, by risk of bias. Interaction p<0.001

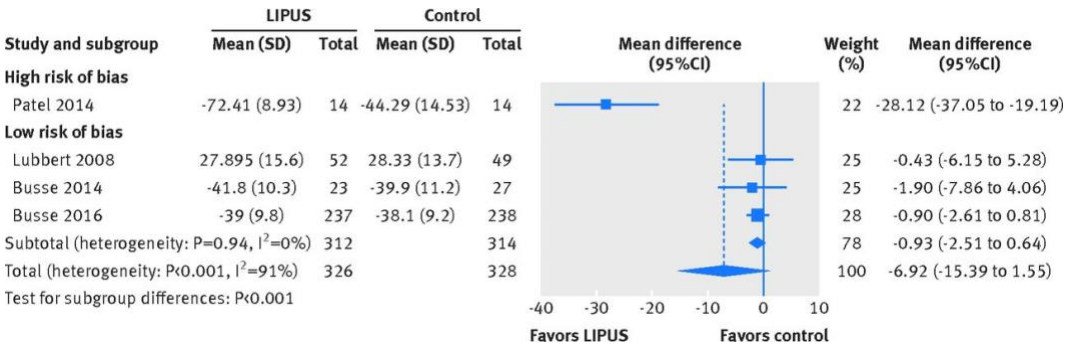


Figure 5: Forest plot for risk ratio for low intensity pulsed ultrasound (LIPUS) compared with control (sham device or no device) of number of subsequent fracture-related operations

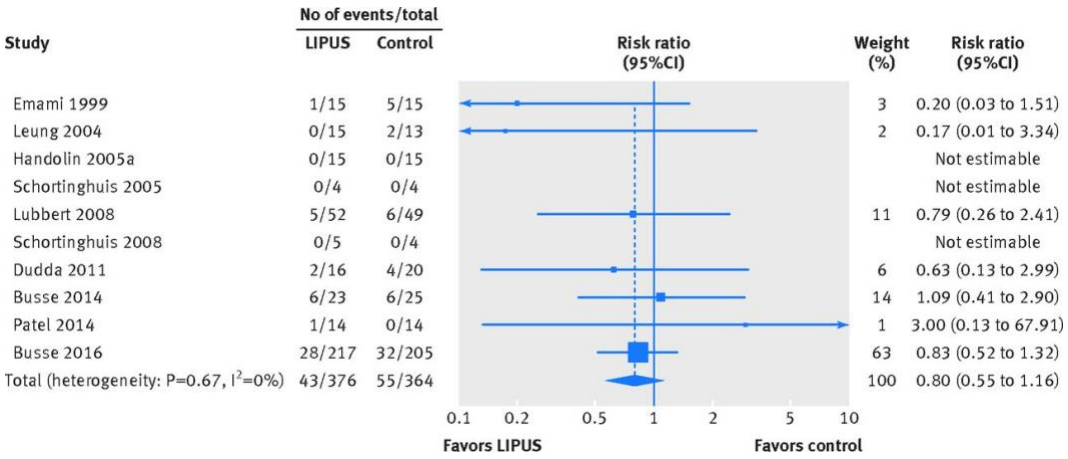


Figure 6: Forest plot for percent difference for low intensity pulsed ultrasound (LIPUS) compared with control (sham device or no device) of days to radiographic healing, by risk of bias. Interaction p<0.001

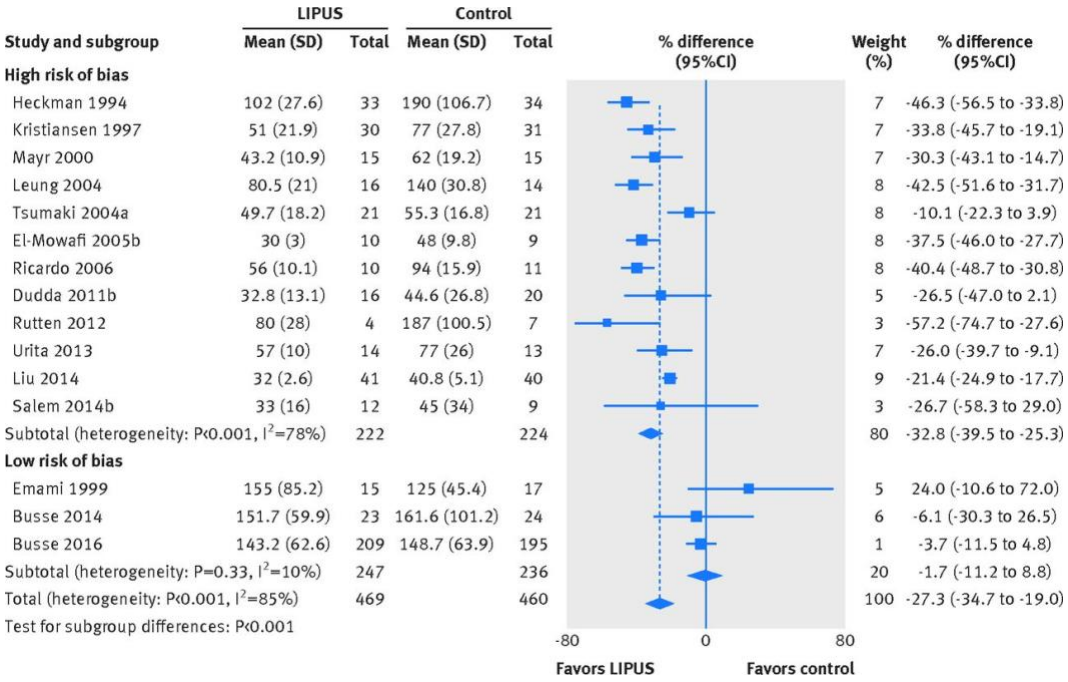
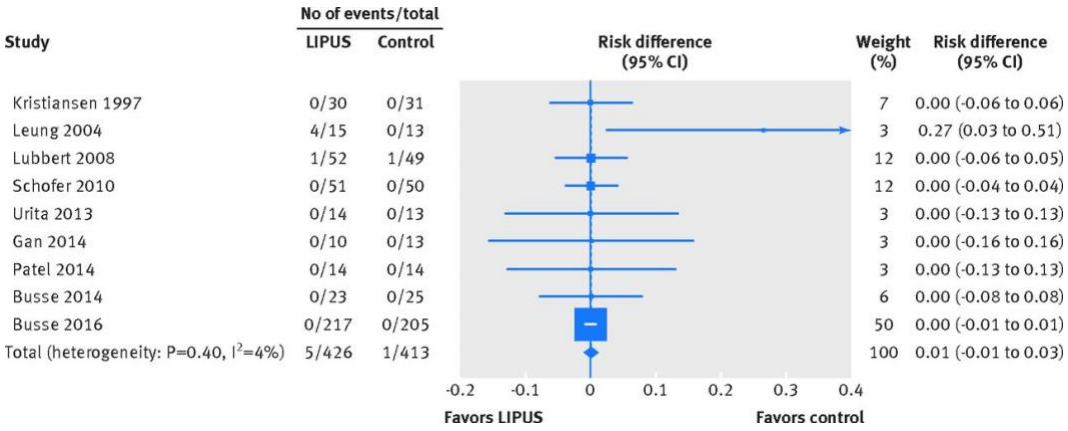


Figure 7: Forest plot for risk difference for low intensity pulsed ultrasound (LIPUS) compared with control (sham device or no device) of ultrasound device related adverse effects



Appendix 1: Literature search strategies

MEDLINE (Ovid)

1 Fracture Healing/
2 Bony Callus/
3 bone remod*.mp.
4 exp Fractures, Bone/
5 fracture*.mp.
6 exp Orthopedic Procedures/
7 or/1-6
8 Ultrasonic Therapy/ or Ultrasonic Waves/
9 LIPUS.mp.
10 8 or 9
11 7 and 10

EMBASE (Ovid)

1 exp fracture healing/
2 callus/
3 bone remod*.mp.
4 exp fracture/
5 (Fracture* and (bone* or osteo* or verteb* or bony or extremity)).mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
6 exp orthopedic surgery/
7 or/1-6
8 exp ultrasound therapy/ or ultrasound.mp. or ultrasonic.mp. or LIPUS.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading]
9 exp echography/
10 8 or 9
11 7 and 10
12 limit 11 to ("therapy (maximizes sensitivity)" or "therapy (maximizes specificity)" or "therapy (best balance of sensitivity and specificity)")

CINAHL (Ebsco)

#Query
S12S11 Limiters - Clinical Queries: Therapy - High Sensitivity, Therapy - High Specificity, Therapy - Best Balance
S11S6 AND S10
S10S8 OR S9
S9(MH "Ultrasonic Therapy")
S8"LIPUS" or "ultrasound"
S7S1 OR S2 OR S3 OR S4 OR S5 OR S6
S6(MH "Orthopedic Surgery+")

S5"fracture*"
S4bone remod*
S3(MH "Bone Remodeling+")
S2(MH "Fracture Healing")
S1(MH "Fractures+")

Cochrane Library (Wiley)
#1MeSH descriptor: [Fracture Healing] explode all trees
#2MeSH descriptor: [Bony Callus] explode all trees
#3MeSH descriptor: [Fractures, Bone] explode all trees
#4bone remod*:ti,ab,kw (Word variations have been searched)
#5fracture*:ti,ab,kw (Word variations have been searched)
#6MeSH descriptor: [Orthopedic Procedures] explode all trees
#7osteotom*
#8#1 or #2 or #3 or #4 or #5 or #6 or #7
#9MeSH descriptor: [Ultrasonic Therapy] explode all trees
#10ultrasound:ti,ab,kw (Word variations have been searched)
#11or #9 or #10
#14#8 and #11 in Trials

PubMed
Search (Therapy/Broad[filter]) AND ((((fracture) AND ultrasound)) AND (((publisher[sb] OR inprocess[sb] OR pubmednotmedline[sb] OR pubstatusaheadofprint)))

Trials registry search in Clinical Trials.gov service of the U.S. National Institutes of Health and World Health Organization International Clinical Trials Registry Platform Search Portal (Search term: low intensity pulsed ultrasound)

Appendix 2: Other functional outcomes

The following outcomes were reported in one study: TRUST Investigators writing group, Busse JW, Bhandari M, Einhorn TA, et al. Re-evaluation of low intensity pulsed ultrasound in treatment of tibial fractures (TRUST): randomized clinical trial. *BMJ (Clinical research ed)* 2016;355:i5351.

Time to return to work, time to event analysis
Hazard ratio 1.11 (95% CI, 0.82 to 1.50) in favour of control

Time to return to leisure activities, time to event analysis
Hazard ratio 1.06 (95% Ci, 0.77 to 1.46) in favour of control

Time to return to ≥80% of pre-injury level of function, time to event analysis
Hazard ratio 1.00 (95% CI, 0.80 to 1.25)

Time to return to full weight bearing, time to event analysis
Hazard ratio time to full weight bearing (0.87; 0.70 to 1.08) in favour of LIPUS

Time to return to household activities, time to event analysis
Percentage difference -1.9% (95%CI, -11.6% to 8.9%) in favour of LIPUS

The following outcomes were reported in two studies: 1) TRUST Investigators writing group, Busse JW, Bhandari M, Einhorn TA, et al. Re-evaluation of low intensity pulsed ultrasound in treatment of tibial fractures (TRUST): randomized clinical trial. *BMJ (Clinical research ed)* 2016;355:i5351, and 2) Lubbert PH, van der Rijt RH, Hoorntje LE, van der Werken C. Low-intensity pulsed ultrasound (LIPUS) in fresh clavicle fractures: a multi-centre double blind randomised controlled trial. *Injury* 2008;39:1444-52.

Figure 1, Forest plot of percent difference for low intensity pulsed ultrasound (LIPUS device) compared with Control (sham device or no device) for days to return to leisure activities

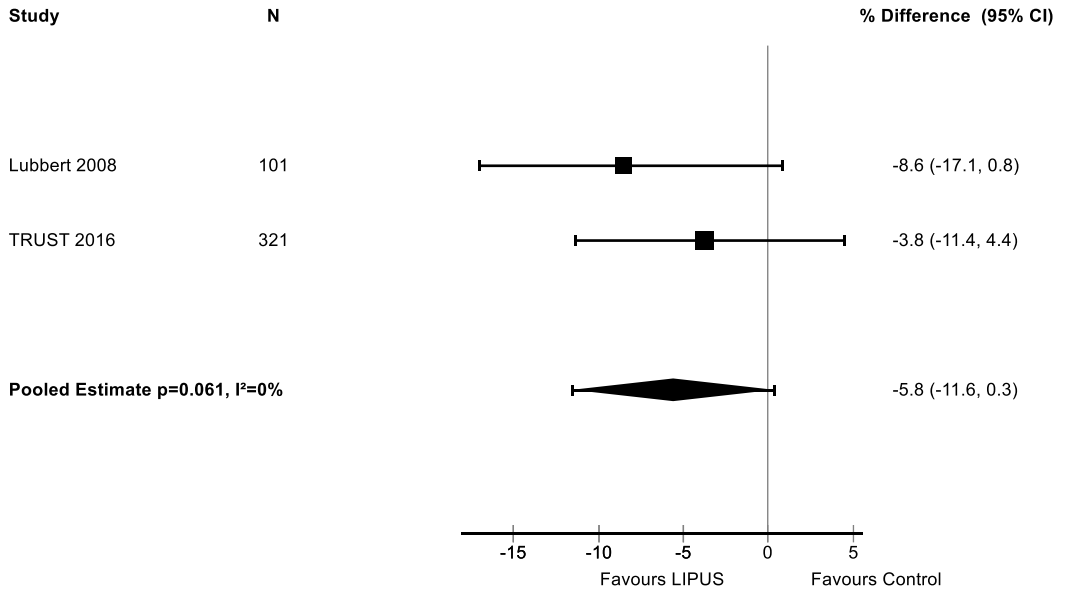
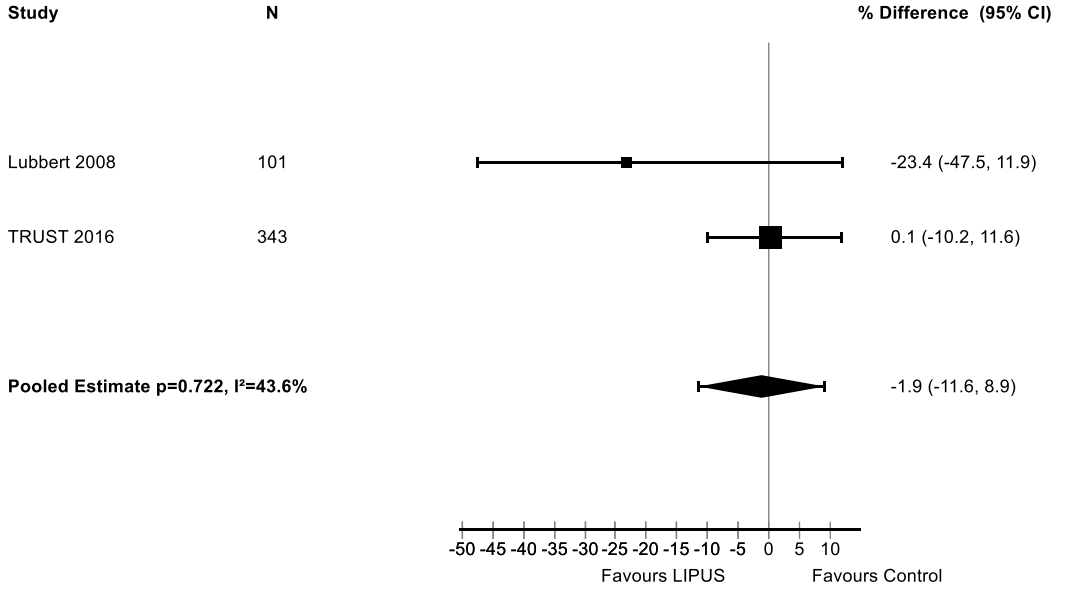


Figure 2, Forest plot of percent difference for low intensity pulsed ultrasound (LIPUS device) compared with Control (sham device or no device) for days to return to household activities



The following outcome was reported in two studies: 1) TRUST Investigators writing group, Busse JW, Bhandari M, Einhorn TA, et al. Re-evaluation of low intensity pulsed ultrasound in treatment of tibial fractures (TRUST): randomized clinical trial. BMJ (Clinical research ed) 2016;355:i5351., and 2) Busse JW, Bhandari M, Einhorn TA, et al. Trial to re-evaluate ultrasound in the treatment of tibial fractures (TRUST): a multicenter randomized pilot study. Trials 2014;15:206.

Multidimensional physical function, continuous, multiple time points

The trials (Busse et al 2014, Busse et al 2014) reported the Short Form 36 physical component summary scores at multiple timepoints. Busse 2016 presented a repeated measurement analysis and found no significant differences in SF-36 scores over time (p=0.30). Busse 2014 provided the plots only.

Figure 3, Busse et al. (2016): Unadjusted repeated measures analyses examining Short Form 36 Physical Component Score (SF-36 PCS) in the Active (LIPUS device) and Sham (sham device) groups found no significant interaction for treatment by time: p=0.30; N ranging from 475 at 6 weeks to 301 at 52 weeks

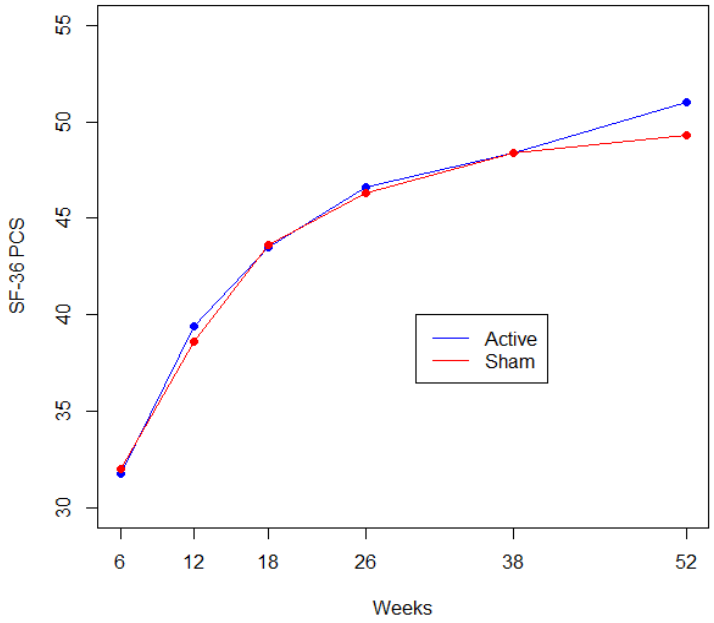
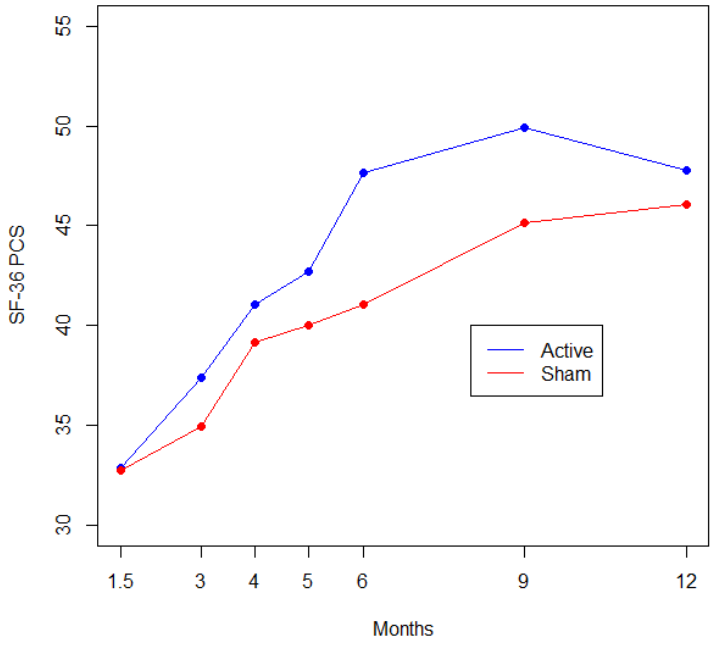


Figure 4, Busse et al. (2014), unpublished material: Unadjusted repeated measures analyses examining Short Form 36 Physical Component Score (SF-36 PCS) in the Active (LIPUS device) and Sham (sham device) groups
N ranging from 50 at 6 weeks to 43 at 1 year.



Appendix 3: Other pain outcomes

The following outcome was reported in two studies: 1) TRUST Investigators writing group, Busse JW, Bhandari M, Einhorn TA, et al. Re-evaluation of low intensity pulsed ultrasound in treatment of tibial fractures (TRUST): randomized clinical trial. BMJ (Clinical research ed) 2016;355:i5351., and 2) Busse JW, Bhandari M, Einhorn TA, et al. Trial to re-evaluate ultrasound in the treatment of tibial fractures (TRUST): a multicenter randomized pilot study. Trials 2014;15:206.

Pain intensity, continuous, multiple time points (unpublished data), subdomain bodily pain of the SF-36 instrument

Figure 1, Busse et al. (2016), unpublished material: Unadjusted repeated measures analyses examining Short Form 36 Bodily Pain (SF-36 Bodily Pain) in the Active (LIPUS device) and Sham (sham device); N ranging from 475 at 6 weeks to 301 at 52 weeks

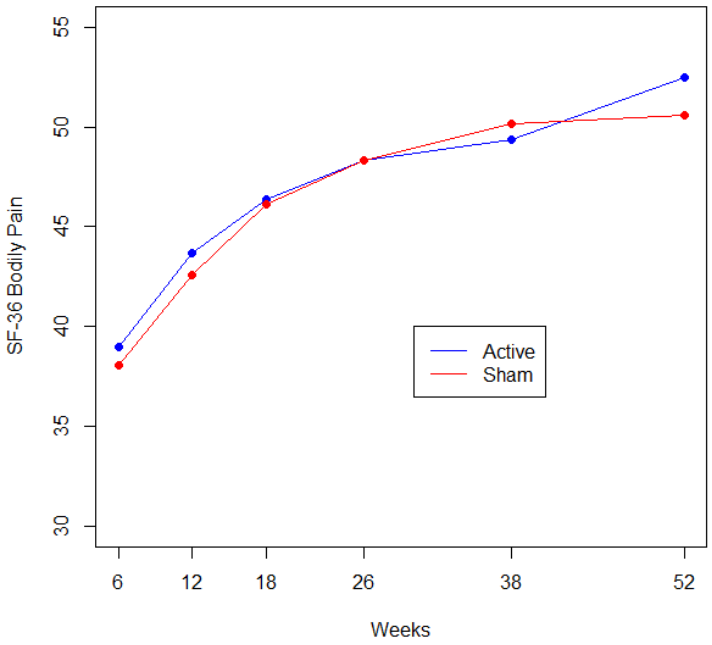
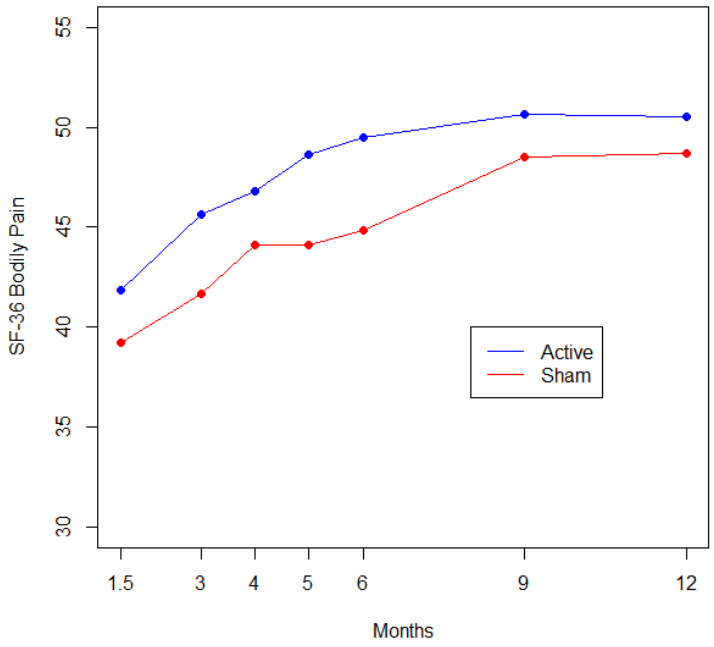
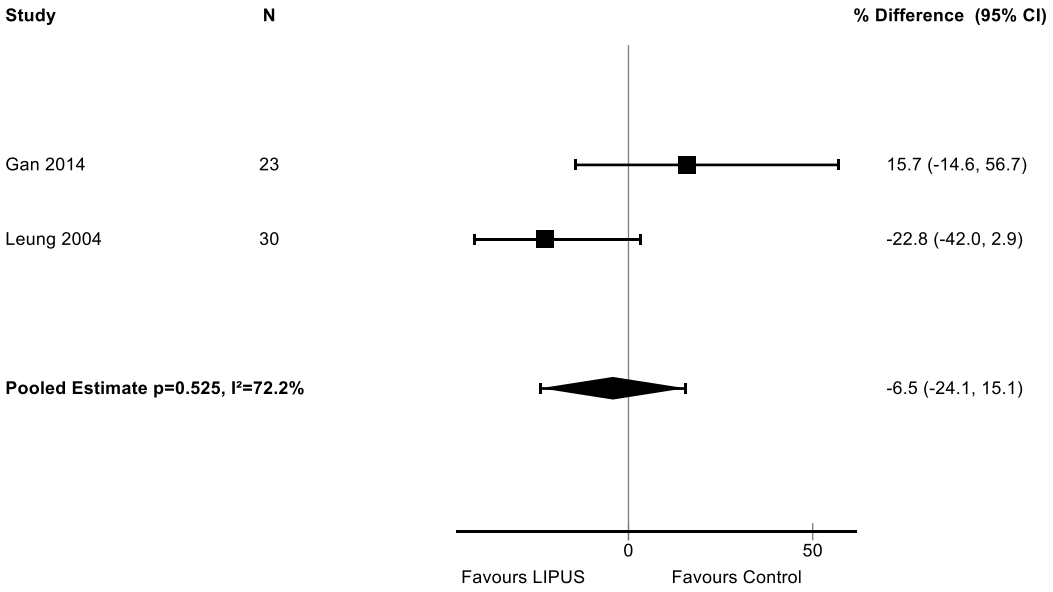


Figure 2, Busse et al. (2014), unpublished material: Unadjusted repeated measures analyses examining Short Form 36 Bodily Pain (SF-36 Bodily Pain) in the Active (LIPUS device) and Sham (sham device); N ranging from 50 at 6 weeks to 43 at 1 year



The following outcome was reported in two studies: 1) Gan TY, Kuah DE, Graham KS, Markson G. Low-intensity pulsed ultrasound in lower limb bone stress injuries: a randomized controlled trial. Clin J Sport Med 2014;24:457-60., and 2) Leung KS, Lee WS, Tsui HF, Liu PP, Cheung WH. Complex tibial fracture outcomes following treatment with low-intensity pulsed ultrasound. Ultrasound Med Biol 2004;30:389-95.

Figure 3, Forest plot percent difference for low intensity pulsed ultrasound (LIPUS device) compared with Control (sham device or no device) for pain duration, number of days with tenderness



Appendix 4: Additional analyses

Figure 1, Forest plot for percent difference for low intensity pulsed ultrasound (LIPUS device) compared with control (sham device or no device) for days to radiographic healing, by clinical subgroups. Interaction p=0.13

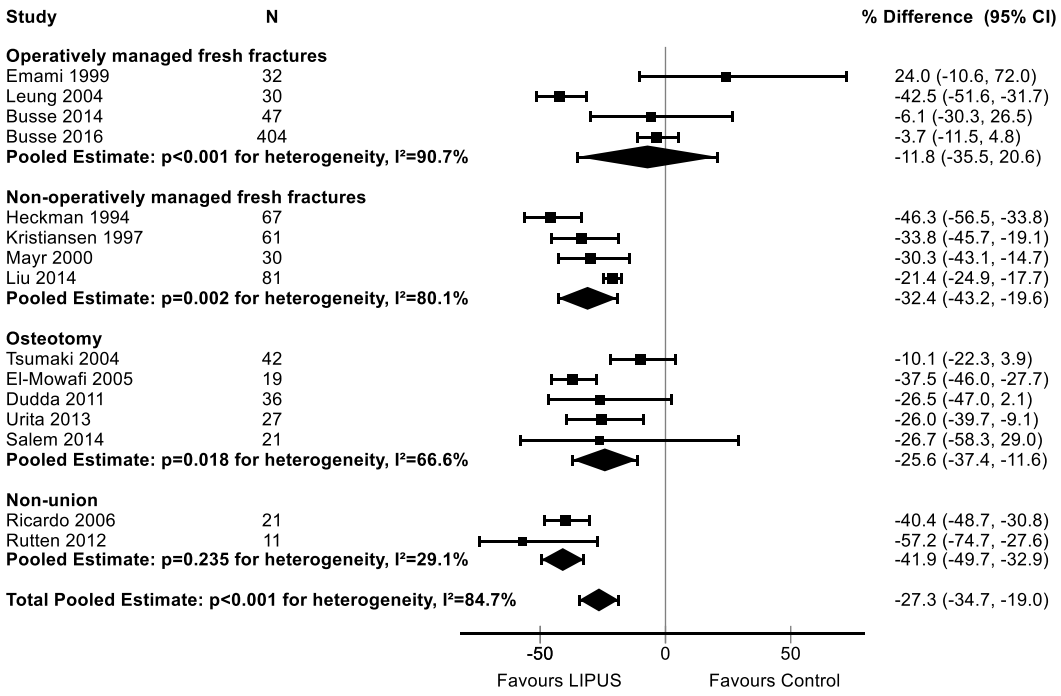


Figure 2, Forest plot for percent difference for low intensity pulsed ultrasound (LIPUS device) compared with control (sham device or no device) for days to radiographic healing, by compliance. Interaction p=0.99

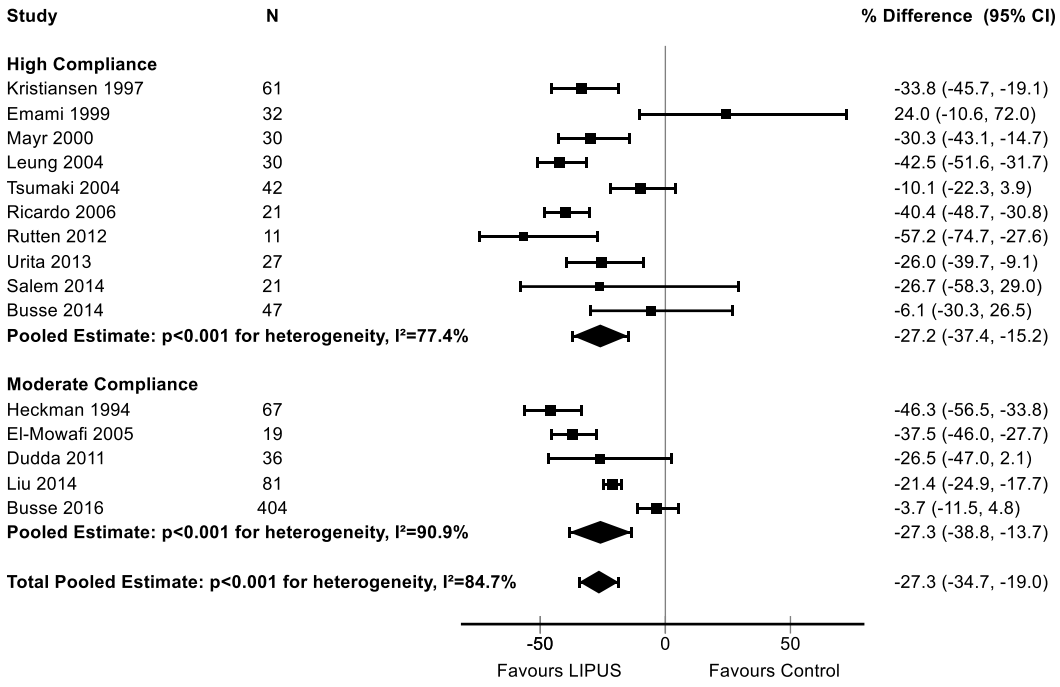


Figure 3, Funnel plot for days to radiographic healing. Egger's linear regression test for asymmetry of the funnel plot: p=0.251

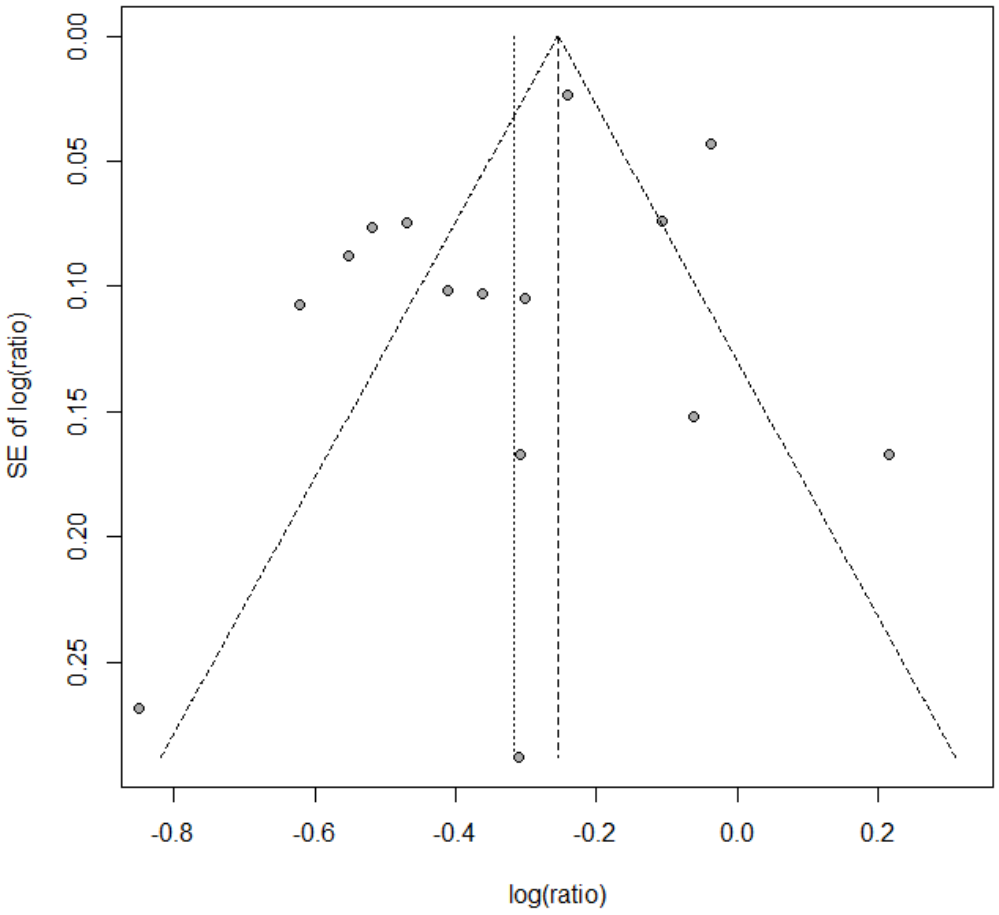
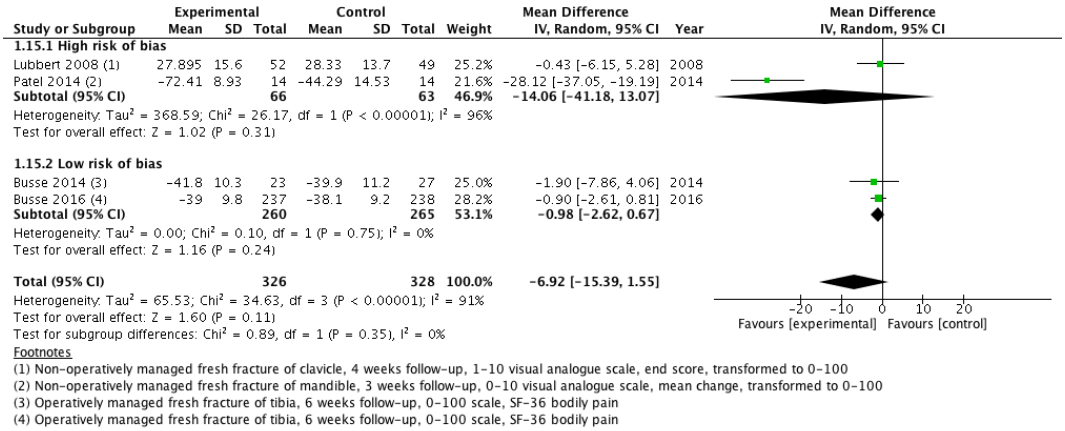


Figure 4, Sensitivity analysis: Forest plot for mean difference for low intensity pulsed ultrasound (LIPUS device) compared with control (sham device or no device) for pain, by risk of bias considering 10% attrition or more representing high risk of bias. Interaction p=0.35



Footnotes

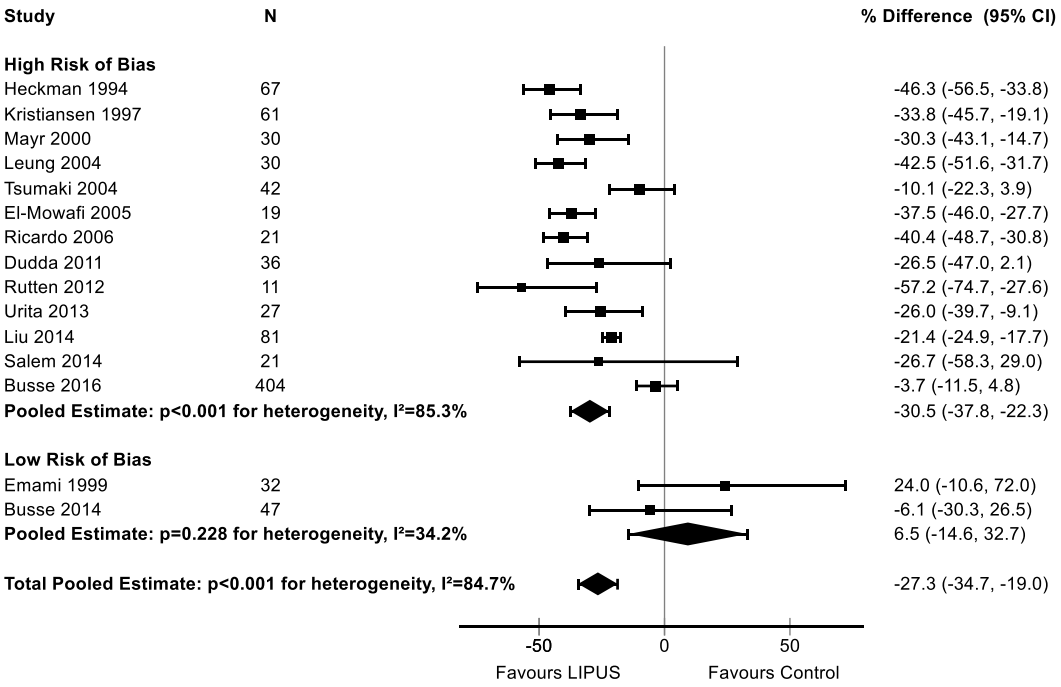
(1) Non-operatively managed fresh fracture of clavicle, 4 weeks follow-up, 1-10 visual analogue scale, end score, transformed to 0-100

(2) Non-operatively managed fresh fracture of mandible, 3 weeks follow-up, 0-10 visual analogue scale, mean change, transformed to 0-100

(3) Operatively managed fresh fracture of tibia, 6 weeks follow-up, 0-100 scale, SF-36 bodily pain

(4) Operatively managed fresh fracture of tibia, 6 weeks follow-up, 0-100 scale, SF-36 bodily pain

Figure 5, Sensitivity analysis: Forest plot for percent difference for low intensity pulsed ultrasound (LIPUS device) compared with control (sham device or no device) for days to radiographic healing, by risk of bias considering 10% attrition or more representing high risk of bias. Interaction p=0.004



Chapter 3: A systematic survey of suggested criteria for assessing the credibility of effect modification in randomized controlled trials or meta-analyses

Currently under review at the Journal of Clinical Epidemiology

Stefan Schandelmaier,[a,b] Yaping Chang,[a] Niveditha Devasenapathy,[c] Tahira Devji,[a] Joey SW 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, ON L8S 4K1, Canada
[b] Basel Institute for Clinical Epidemiology and Biostatistics, Department of Clinical Research, 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, ON L8S 4K1, Canada
[f] Department of Medicine, University of Toronto, 190 Elizabeth Street, R. Fraser Elliott Building, 3-805, Toronto, ON M5G 2C4, Canada
[g] Center for Evidence-based Chinese Medicine, Beijing University of Chinese Medicine, 11 Bei San Huan Dong Lu, Chaoyang District, 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, ON L8N 4A6, Hamilton, Canada
[j] Department of Medicine, McMaster University, 1200 Main Street West, Hamilton, ON L8S 4L8, Hamilton, Canada

Abstract

Objective: To systematically survey the methodological literature and collect suggested criteria for assessing the credibility of effect modification and associated rationales.

Study design: We searched MEDLINE, Embase, textbook chapters 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 pre-specification 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 performance of criteria in practice.

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 is known:

Key differences from 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, 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.

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 upon may be to the patient's detriment[1].

The methodological literature has widely acknowledged the challenges of dealing with putative effect modification. 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.[34] 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.

Methods

Eligibility criteria

We included publications (journal articles, reports, text book 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, i.e. 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 considered any synonym or paraphrase 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 such as "review" or "summary", or exclusive referencing of other publications.

Search strategy

In collaboration with an experienced medical librarian (NB), 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 (StS) 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.

Data abstraction

We designed and pre-tested 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 below). After participating in an initial calibration exercise, pairs of reviewers independently abstracted data, resolving disagreements through discussion. To ensure consistency of judgments, StS was a member of all reviewer pairs (i.e. StS and one of YC, ND, JK, LEC, YL, AA, TD, or YZ).

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 papers 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 qualitative practice test to find out whether users find the proposed checklist useful, comprehensive, relevant, and easy-to-use, 4) a reliability study (i.e. a quantitative practice test

to find out whether the criteria actually helps users differentiate between more or less credible subgroup effects), or 5) other formal methods for instrument development or testing.

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.

Results

Search results

We screened 2117 records or journal publications and tables of contents of 151 textbooks, assessed 557 publications in full text, and finally included 150 publications (Figure 1). The dates of publication spanned four decades. Publications were mostly journal articles (n=130), focused on individual RCTs (n=97) and provided up to 15 criteria (Table 1).

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 non-significant 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 amongst 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 twelve 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).

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).

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); pre-specified 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 pre-specifying analytic details) and on the analysis level (e.g. applying a formal test of interaction, accounting for multiplicity, shrinking estimates towards 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, pre-specified 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.[10-32]

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,[33] performing exploratory subgroup analyses,[34] modelling continuous effect modifiers[35-37], or addressing the correlation between subgrouping variables.[38] 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 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.[39] 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. 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 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. clinicians, systematic review authors, guideline developers, journal editors, policy makers). Determining the feasibility, acceptability and reliability of any instrument suggested for wide use would therefore be crucial.

Acknowledgements
We thank Kuebra Oezoglu for gathering full text articles.

Funding sources
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.

References

1.Rothwell PM. **Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation.** Lancet. 2005;365(9454):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 medical research methodology. 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 medical research methodology. 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; Methods Research Paper AHRQ Publication No 10-EHC070-EF Available at http://effectivehealthcareahrqgov/. 2010.
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. pcori.org; 2012.
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 medical research methodology. 2017;17(1):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 medical research methodology. 2013;13:106.
8.VanderWeele TJ, Knol MJ. **Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions.** Annals of internal medicine. 2011;154(10):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 technology assessment (Winchester, England). 2001;5(33):1-56.
10.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.
11.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.
12.Gabler NBD, N.; Liao, D.; Elmore, J. G.; Ganiats, T. G.; Kravitz, R. L. **Dealing with heterogeneity of treatment effects: is the literature up to the challenge?** Trials. 2009;10:43.
13.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.** Journal of clinical epidemiology. 2016;79:3-9.
14.Simmonds M, Stewart G, Stewart L. **A decade of individual participant data meta-analyses: A review of current practice.** Contemporary clinical trials. 2015;45(Pt A):76-83.

15.Zhang S, Liang F, Li W, Hu X. **Subgroup Analyses in Reporting of Phase III Clinical Trials in Solid Tumors.** Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2015;33(15):1697-702.
16.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(6):e0128804.
17.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.** European journal of cancer. 2015;51(18):2732-9.
18.Mistry DP, S.; Hee, S. W.; 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.
19.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.
20.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.
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(1):283-301.
22.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.** International journal of technology assessment in health care. 2008;24(3):358-61.
23.Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. **Statistics in medicine--reporting of subgroup analyses in clinical trials.** The New England journal of medicine. 2007;357(21):2189-94.
24.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.** Journal of clinical epidemiology. 2007;60(10):1002-9.
25.Aulakh AK, Anand SS. **Sex and gender subgroup analyses of randomized trials.** Women's health issues : official publication of the Jacobs Institute of Women's Health. 2007;17(6):342-50.
26.Patsopoulos NA, Tatsioni A, Ioannidis JP. **Claims of sex differences: an empirical assessment in genetic associations.** Jama. 2007;298(8):880-93.
27.Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. **Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading?** American heart journal. 2006;151(2):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.** Clinical orthopaedics and related research. 2006;447:247-51.
29.Higgins JT, S.; Deeks, J.; Altman, D. **Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice.** Journal of health services research & policy. 2002;7(1):51-61.

30.Pocock SJ, Assmann SE, Enos LE, Kasten LE. **Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems.** Statistics in medicine. 2002;21(19):2917-30.
31.Assmann SF, Pocock SJ, Enos LE, Kasten LE. **Subgroup analysis and other (mis)uses of baseline data in clinical trials.** Lancet. 2000;355(9209):1064-9.
32.Parker ABN, C. D. **Subgroups, treatment effects, and baseline risks: some lessons from major cardiovascular trials.** American heart journal. 2000;139(6):952-61.
33.Alosh M, Huque MF. **Multiplicity considerations for subgroup analysis subject to consistency constraint.** Biometrical journal Biometrische Zeitschrift. 2013;55(3):444-62.
34.Lipkovich I, Dmitrienko A, B. R. D' Agostino S. **Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials.** Statistics in medicine. 2017;36(1):136-96.
35.Royston PS, W. **A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials.** Stat Med. 2004;23(16):2509-25.
36.Royston PS, W. **Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis.** Stat Med. 2013;32(22):3788-803.
37.Royston P, Sauerbrei W. **Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis.** Statistics in medicine. 2014;33(27):4695-708.
38.Varadhan R, Wang SJ. **Standardization for subgroup analysis in randomized controlled trials.** Journal of biopharmaceutical statistics. 2014;24(1):154-67.
39.European Medicines Agency. Guideline on the investigation of subgroups in confirmatory clinical trials (draft). London, UK2014.
40.Burke JF, Sussman JB, Kent DM, Hayward RA. **Three simple rules to ensure reasonably credible subgroup analyses.** Bmj. 2015;351:h5651.
41.Koch A, Framke T. **Reliably basing conclusions on subgroups of randomized clinical trials.** Journal of biopharmaceutical statistics. 2014;24(1):42-57.
42.Wang SJ, Hung HM. **A regulatory perspective on essential considerations in design and analysis of subgroups when correctly classified.** Journal of biopharmaceutical statistics. 2014;24(1):19-41.
43.Desai M, Pieper KS, Mahaffey K. **Challenges and solutions to pre- and post-randomization subgroup analyses.** Current cardiology reports. 2014;16(10):531.
44.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(4):405-11.
45.Paget MA, Chuang-Stein C, Fletcher C, Reid C. **Subgroup analyses of clinical effectiveness to support health technology assessments.** Pharmaceutical statistics. 2011;10(6):532-8.
46.Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Higgins JP, Green SB, editors: The Cochrane Collaboration; 2011.
47.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.
48.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.
49.Dijkman B, Kooistra B, Bhandari M, Evidence-Based Surgery Working G. **How to work with a subgroup analysis.** Canadian journal of surgery Journal canadien de chirurgie. 2009;52(6):515-22.

50.Fletcher J. **Subgroup analyses: how to avoid being misled.** Bmj. 2007;335(7610):96-7.
51.Grouin JM, Coste M, Lewis J. **Subgroup analyses in randomized clinical trials: statistical and regulatory issues.** Journal of biopharmaceutical statistics. 2005;15(5):869-82.
52.Cook DI, Gebski VJ, Keech AC. **Subgroup analysis in clinical trials.** The Medical journal of Australia. 2004;180(6):289-91.
53.Moreira ED, Susser E. **Guidelines on how to assess the validity of results presented in subgroup analysis of clinical trials.** Revista do Hospital das Clinicas. 2002;57(2):83-8.
54.Oxman AD, Guyatt GH. **A consumer's guide to subgroup analyses.** Annals of internal medicine. 1992;116(1):78-84.
55.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(1):93-8.
56.Weiss NS. **Subgroup-specific associations in the face of overall null results: should we rush in or fear to tread?** Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(6):1297-9.
57.Thompson SG, Higgins JP. **How should meta-regression analyses be undertaken and interpreted?** Statistics in medicine. 2002;21(11):1559-73.
58.Sleight P. **Debate: Subgroup analyses in clinical trials: fun to look at – but don't believe them!** Current controlled trials in cardiovascular medicine. 2000;1(1):25-7.
59.Simon R. **Patient subsets and variation in therapeutic efficacy.** British journal of clinical pharmacology. 1982;14(4):473-82.
60.Bulpitt CJ. **Subgroup analysis.** Lancet. 1988;2(8601):31-4.
61.Lagakos SW. **The challenge of subgroup analyses--reporting without distorting.** The New England journal of medicine. 2006;354(16):1667-9.
62.Jones HE, Ohlssen DI, Neuenschwander B, Racine A, Branson M. **Bayesian models for subgroup analysis in clinical trials.** Clinical trials (London, England). 2011;8(2):129-43.
63.Gail M, Simon R. **Testing for qualitative interactions between treatment effects and patient subsets.** Biometrics. 1985;41(2):361-72.
64.Berlin JA. **Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies.** American journal of epidemiology. 1995;142(4):383-7.
65.Laud PW, Sivaganesan S, Müller P. Subgroup Analysis. In: Damien P, Dellaportas P, Polson NG, Stephens DA, editors. Bayesian Theory and Applications. Oxford: Oxford University Press; 2013.
66.Grady D, Cummings SR, Hulley SB. Chapter 11: Alternative trial designs and implementation issues. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB, editors. Designing Clinical Research. 3 ed. Philadelphia: LIPPINCOTT WILLIAMS & WILKINS,; 2007.
67.Moye LA. Chater 21: The multiple comparison issue in health care research. In: Rao CR, Miller JP, Rao DC, editors. Handbok of statistics: epidemiology and medical statistics. 1 ed. Amsterdam: Elsevier; 2008.
68.Borenstein M, Hedges L, Higgins JP, Rothstein H. Introduction to meta-analysis. 1 ed. Chichester: John Wiley & Sons; 2009.
69.Dahabreh IJ, Trikalinos TA, Kent DM, Schmid CH. Heterogeneity of Treatment Effects. In: Gatsonis C, Morton SC, editors. Methods in Comparative Effectiveness Research. Boca Raton: CRC Press; 2017.

70.Bulpitt CJ. Randomized Controlled Clinical Trials. 2 ed: Springer Science and Business Media, LLC; 1996.
71.Starr JR, McKnight B. **Assessing interaction in case-control studies: type I errors when using both additive and multiplicative scales.** Epidemiology. 2004;15(4):422-7.
72.The National Institute for Health and Care Excellence (NICE). Guide to the methods of technology appraisal 2013: Process and methods2013.
73.Cui L, Hung HM, Wang SJ, Tsong Y. **Issues related to subgroup analysis in clinical trials.** Journal of biopharmaceutical statistics. 2002;12(3):347-58.
74.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. 2014;39(7):618-29.
75.Fleming TR. **Interpretation of subgroup analyses in clinical trials.** Therapeutic Innovation & Regulatory Science 1995;29(1):1681S-7S.
76.Grobbee DE, Hoes AW. Chapter 3: Etiological Research. Clinical Epidemiology: Principles, Methods, and Applications for Clinical Research. 2nd ed. Burlington, MA: Jones & Bartlett Learning; 2015.
77.Meinert CL. Chapter 17: An Insider's Guide to Clinical Trials. An Insider's Guide to Clinical Trials. 1 ed. Oxford: Oxford University Press; 2011.
78.Beaujean AA. Mediation, Moderation, and the Study of Individual Differences. In: Osborne J, editor. Best Practices in Quantitative Methods. Thousand Oaks: SAGE publications; 2008.
79.Weiss CO, Segal JB, Varadhan R. **Assessing the applicability of trial evidence to a target sample in the presence of heterogeneity of treatment effect.** Pharmacoepidemiology and drug safety. 2012;21 Suppl 2:121-9.
80.Thompson SG, Higgins JP. **Treating individuals 4: can meta-analysis help target interventions at individuals most likely to benefit?** Lancet. 2005;365(9456):341-6.
81.Kamper SJ, Maher CG, Hancock MJ, Koes BW, Croft PR, Hay E. **Treatment-based subgroups of low back pain: a guide to appraisal of research studies and a summary of current evidence.** Best practice & research Clinical rheumatology. 2010;24(2):181-91.
82.Feinstein AR. **The problem of cogent subgroups: a clinicostatistical tragedy.** Journal of clinical epidemiology. 1998;51(4):297-9.
83.Schulz KF, Grimes DA. **Multiplicity in randomised trials II: subgroup and interim analyses.** Lancet. 2005;365(9471):1657-61.
84.Altman DG. **Clinical trials: subgroup analyses in randomized trials--more rigour needed.** Nat Rev Clin Oncol. 2015;12(9):506-7.
85.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.** International journal of epidemiology. 2016;45(6):2184-93.
86.Russek-Cohen E, Simon RM. **Qualitative interactions in multifactor studies.** Biometrics. 1993;49(2):467-77.
87.Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. **Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies.** Statistics in medicine. 2000;19(24):3325-36.
88.VanderWeele TJ, Knol MJ. **A Tutorial on Interaction.** Epidemiologic Methods. 2014;3:33-72.

89.Alosh M, Huque MF, Bretz F, D'Agostino RB, Sr. **Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials.** Statistics in medicine. 2017;36(8):1334-60.
90.Thompson SG. **Why sources of heterogeneity in meta-analysis should be investigated.** Bmj. 1994;309(6965):1351-5.
91.Schmid CH, Stark PC, Berlin JA, Landais P, Lau J. **Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors.** Journal of clinical epidemiology. 2004;57(7):683-97.
92.Lambert PC, Sutton AJ, Abrams KR, Jones DR. **A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis.** Journal of clinical epidemiology. 2002;55(1):86-94.
93.Davey Smith G, Egger M, Phillips AN. **Meta-analysis. Beyond the grand mean?** Bmj. 1997;315(7122):1610-4.
94.Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI, Anti-Lymphocyte Antibody Induction Therapy Study G. **Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head.** Statistics in medicine. 2002;21(3):371-87.
95.Thompson SG, Sharp SJ. **Explaining heterogeneity in meta-analysis: a comparison of methods.** Statistics in medicine. 1999;18(20):2693-708.
96.Smith CT, Williamson PR, Marson AG. **An overview of methods and empirical comparison of aggregate data and individual patient data results for investigating heterogeneity in meta-analysis of time-to-event outcomes.** Journal of evaluation in clinical practice. 2005;11(5):468-78.
97.Simmonds MC, Higgins JP. **Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data.** Statistics in medicine. 2007;26(15):2982-99.
98.Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. **Prognosis research strategy (PROGRESS) 4: stratified medicine research.** Bmj. 2013;346:e5793.
99.Donegan S, Williamson P, D'Alessandro U, Tudur Smith C. **Assessing the consistency assumption by exploring treatment by covariate interactions in mixed treatment comparison meta-analysis: individual patient-level covariates versus aggregate trial-level covariates.** Statistics in medicine. 2012;31(29):3840-57.
100.Egger M, Smith GD, Altman DG. Systematic reviews in health care: meta-analysis in context. 2 ed. London: BMJ Publishing Group; 2001.
101.Hua. **One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information.** Statistics in medicine. 2017.
102.Davey Smith G, Egger M. Going beyond the grand mean: subgroup analysis in meta-analysis of randommized trials. In: Egger M, Davey Smith G, Altman DG, editors. Systematic reviews in helath care: meta-analysis in context. 2 ed. London: BMJ; 2001.
103.Stephens R. **The dangers of subgroup analysis.** The Lancet Oncology. 2001;2(1):9.
104.Matsuyama Y, Morita S. **Estimation of the average causal effect among subgroups defined by post-treatment variables.** Clinical trials (London, England). 2006;3(1):1-9.
105.Hirji KF, Fagerland MW. **Outcome based subgroup analysis: a neglected concern.** Trials. 2009;10:33.

106.Cuzick J. **The assessment of subgroups in clinical trials.** Experientia Supplementum. 1982;41:224-35.
107.Wang R, Ware JH. **Detecting moderator effects using subgroup analyses.** Prevention science : the official journal of the Society for Prevention Research. 2013;14(2):111-20.
108.Korn EL, Othus M, Chen T, Freidlin B. **Assessing treatment efficacy in the subset of responders in a randomized clinical trial.** Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2017;28(7):1640-7.
109.Peto R. **Current misconception 3: that subgroup-specific trial mortality results often provide a good basis for individualising patient care.** British journal of cancer. 2011;104(7):1057-8.
110.Higgins JP, Thompson SG. **Controlling the risk of spurious findings from meta-regression.** Statistics in medicine. 2004;23(11):1663-82.
111.Biesheuvel EH, Hothorn LA. **Protocol designed subgroup analyses in multiarmed clinical trials: multiplicity aspects.** Journal of biopharmaceutical statistics. 2003;13(4):663-73.
112.Alosh M, Huque MF. **A flexible strategy for testing subgroups and overall population.** Statistics in medicine. 2009;28(1):3-23.
113.Kerstenbaum B. Chapter 7: Randomized Trials. In: Adeney KL, Weiss NS, editors. Epidemiology and Biostatistics: An Introduction to Clinical Research. Seatllle, WA: Springer Science+Business Media, LLC; 2009.
114.Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. **CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials.** Bmj. 2010;340:c869.
115.Schmidt AF, Klungel OH, Nielen M, de Boer A, Groenwold RH, Hoes AW. **Tailoring treatments using treatment effect modification.** Pharmacoepidemiology and drug safety. 2016;25(4):355-62.
116.Taubman B. **Clinical trial of the treatment of colic by modification of parent-infant interaction.** Pediatrics. 1984;74(6):998-1003.
117.Kaiser LD. **Stratification of randomization is not required for a pre-specified subgroup analysis.** Pharmaceutical statistics. 2013;12(1):43-7.
118.Song Y, Chi GY. **A method for testing a prespecified subgroup in clinical trials.** Statistics in medicine. 2007;26(19):3535-49.
119.Eisner MD. **The challenge of subgroup analyses.** The New England journal of medicine. 2006;355(2):211; author reply -2.
120.Kaiser LD. **Inefficiency of randomization methods that balance on stratum margins and improvements with permuted blocks and a sequential method.** Statistics in medicine. 2012;31(16):1699-706.
121.Sharp SJ, Thompson SG, Altman DG. **The relation between treatment benefit and underlying risk in meta-analysis.** Bmj. 1996;313(7059):735-8.
122.Senn S. **Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials.** Statistics in medicine. 1994;13(3):293-6.
123.Furberg CD, Byington RP. **What do subgroup analyses reveal about differential response to beta-blocker therapy? The Beta-Blocker Heart Attack Trial experience.** Circulation. 1983;67(6 Pt 2):I98-101.

124.Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. **Estimates of subgroup treatment effects in overall nonsignificant trials: To what extent should we believe in them?** Pharmaceutical statistics. 2017;16(4):280-95.
125.Rubio-Aparicio M, Sanchez-Meca J, Lopez-Lopez JA, Botella J, Marin-Martinez F. **Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances.** The British journal of mathematical and statistical psychology. 2017;70(3):439-56.
126.Dissemination CfRa. Systematic Reviews: CRD's guidance for undertaking reviews in health care. York: Centre for Reviews and Dissemination; 2009.
127.Gelber RD, Goldhirsch A. **Interpretation of results from subset analyses within overviews of randomized clinical trials.** Statistics in medicine. 1987;6(3):371-8.
128.VanderWeele TJ. Explanation in causal inference. Methods for mediation and interaction. 1 ed. New York: Oxford University Press; 2015.
129.Pearce N, Greenland S. Confounding and Interaction In: Ahrens W, Pigeot I, editors. Handbook of Epidemiology. 2 ed. New York: Springer Science + Business Media; 2014.
130.Cuzick J. **Forest plots and the interpretation of subgroups.** Lancet. 2005;365(9467):1308.
131.Matthews JN, Altman DG. **Statistics notes. Interaction 2: Compare effect sizes not P values.** Bmj. 1996;313(7060):808.
132.Rockette HE, Caplan RJ. **Strategies for subgroup analysis in clinical trials.** Recent results in cancer research Fortschritte der Krebsforschung Progres dans les recherches sur le cancer. 1988;111:49-54.
133.Scott PEC, G. **Interpretation of subgroup analyses in medical device clinical trials.** Drug Inf J. 1998;32(1):213-20.
134.Royston P, Sauerbrei W. **A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials.** Statistics in medicine. 2004;23(16):2509-25.
135.Varadhan R, Wang SJ. **Treatment effect heterogeneity for univariate subgroups in clinical trials: Shrinkage, standardization, or else.** Biometrical journal Biometrische Zeitschrift. 2016;58(1):133-53.
136.Berry DA. **Subgroup analyses.** Biometrics. 1990;46(4):1227-30.
137.Song F, Bachmann MO. **Cumulative subgroup analysis to reduce waste in clinical research for individualised medicine.** BMC medicine. 2016;14(1):197.
138.Koch GG, Schwartz TA. **An overview of statistical planning to address subgroups in confirmatory clinical trials.** Journal of biopharmaceutical statistics. 2014;24(1):72-93.
139.Lipkovich I, Dmitrienko A, Muysers C, Ratitch B. **Multiplicity issues in exploratory subgroup analysis.** Journal of biopharmaceutical statistics. 2018;28(1):63-81.
140.Counsell CE, Clarke MJ, Slattery J, Sandercock PA. **The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis?** Bmj. 1994;309(6970):1677-81.
141.Varadhan R, Segal JB, Boyd CM, Wu AW, Weiss CO. **A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research.** Journal of clinical epidemiology. 2013;66(8):818-25.
142.Groenwold RH, Donders AR, van der Heijden GJ, Hoes AW, Rovers MM. **Confounding of subgroup analyses in randomized data.** Archives of internal medicine. 2009;169(16):1532-4.

143.Glasziou PP, Sanders SL. **Investigating causes of heterogeneity in systematic reviews.** Statistics in medicine. 2002;21(11):1503-11.
144.Moye LA, Deswal A. **Trials within trials: confirmatory subgroup analyses in controlled clinical experiments.** Controlled clinical trials. 2001;22(6):605-19.
145.Greenland S. **Interactions in epidemiology: relevance, identification, and estimation.** Epidemiology. 2009;20(1):14-7.
146.Rothman KJ, Greenland S, Walker AM. **Concepts of interaction.** American journal of epidemiology. 1980;112(4):467-70.
147.Keene ON, Garrett AD. **Subgroups: time to go back to basic statistical principles?** Journal of biopharmaceutical statistics. 2014;24(1):58-71.
148.Royston P, Sauerbrei W. **Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis.** Statistics in medicine. 2013;32(22):3788-803.
149.Royston P, Sauerbrei W. **Interactions between treatment and continuous covariates: a step toward individualizing therapy.** Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2008;26(9):1397-9.
150.Wu AD, Zumbo BD. **Understanding and Using Mediators and Moderators.** Social Indicators Research. 2007;87(3):367-92.
151.Gilthorpe MS, Clayton DG. Statistical Interactions and Gene-Environment Joint Effects. In: Tu YK, Greenwood DC, editors. Modern methods for Epidemiology. 1 ed. Dordrecht: Springer; 2012.
152.Borenstein M, Higgins JP. **Meta-analysis and subgroups.** Prevention science : the official journal of the Society for Prevention Research. 2013;14(2):134-43.
153.Donner A. **A Bayesian approach to the interpretation of subgroup results in clinical trials.** Journal of chronic diseases. 1982;35(6):429-35.
154.Sivaganesan S, Laud PW, Muller P. **A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme.** Statistics in medicine. 2011;30(4):312-23.
155.Clarke M, Halsey J. **DICE 2: a further investigation of the effects of chance in life, death and subgroup analyses.** International journal of clinical practice. 2001;55(4):240-2.
156.Yusuf S, Wittes J. **Interpreting Geographic Variations in Results of Randomized, Controlled Trials.** The New England journal of medicine. 2016;375(23):2263-71.
157.Altman DG, Matthews JN. **Statistics notes. Interaction 1: Heterogeneity of effects.** Bmj. 1996;313(7055):486.
158.Mullins CD. **Subgroup analysis versus post-hoc analysis.** Clinical therapeutics. 2001;23(7):1060.
159.Byar DP. **Assessing apparent treatment--covariate interactions in randomized clinical trials.** Statistics in medicine. 1985;4(3):255-63.
160.Altman DG. **Within Trial Variation—A False Trail?** Journal of clinical epidemiology. 1998;51(4):301-3.
161.Collins R, Gray R, Godwin J, Peto R. **Avoidance of large biases and large random errors in the assessment of moderate treatment effects: the need for systematic overviews.** Statistics in medicine. 1987;6(3):245-54.
162.Oxman AD. **Subgroup analyses.** Bmj. 2012;344:e2022.

163.Goldberger JJ, Buxton AE. **Personalized medicine vs guideline-based medicine.** Jama. 2013;309(24):2559-60.
164.Abrahamowicz M, Beauchamp ME, Fournier P, Dumont A. **Evidence of subgroup-specific treatment effect in the absence of an overall effect: is there really a contradiction?** Pharmacoepidemiology and drug safety. 2013;22(11):1178-88.
165.Scales CD, Jr., Canfield SE. **Advanced topics in evidence-based urological oncology: using results of a subgroup analysis.** Urologic oncology. 2011;29(4):462-6.
166.Hahn S, Garner P, Williamson P. **Are systematic reviews taking heterogeneity into account? An analysis from the Infectious Diseases Module of the Cochrane Library.** Journal of evaluation in clinical practice. 2000;6(2):231-3.

Figure 1: Study selection flow chart

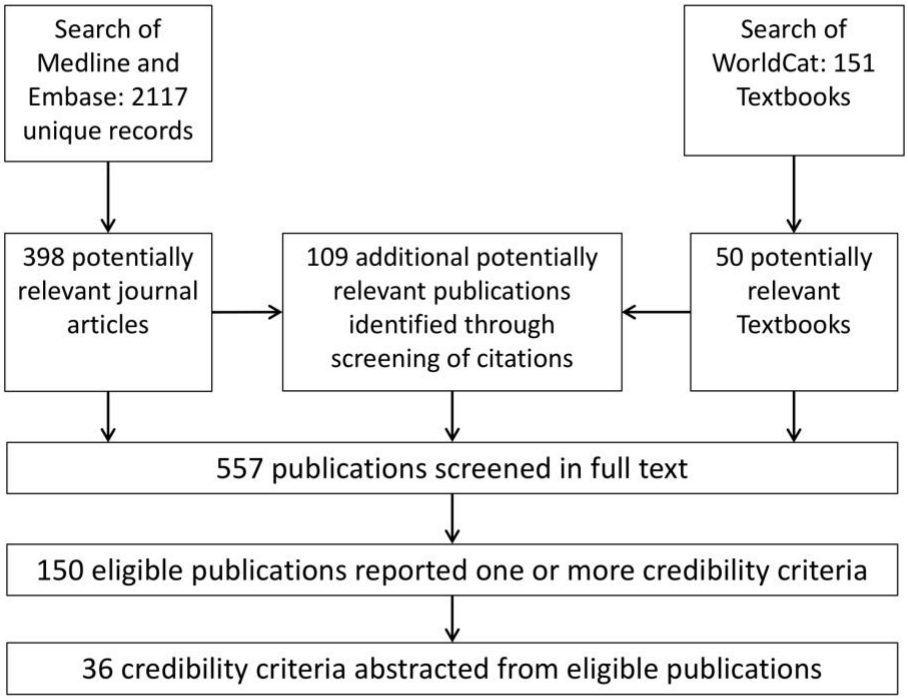


Table 1: Characteristics of the 150 included publications

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

[a] Cochrane collaboration (REF), Centre 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 to included publications)

	Criterion	n
Design characteristics	**Pre-specification of analytic details [a]:** Credibility is higher if analytic details such as cut points, time-points, or statistical methods have been pre-specified prior to the analysis	39
	Small number of candidate effect modifiers: Credibility is higher if only a small number of effect modifiers have been tested	38
	Within versus 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
	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
	Power: Credibility increases with the power to detect the effect modifier	23
	Stratified randomization: Credibility is higher if the effect modifier was a stratification variable at randomization	15
	Regression on control group risk: [b] Credibility is lower if the effect modifier in a meta-analysis is defined by the risk of an outcome in the control group	6
	Primary outcome: Credibility is higher for effect modifiers claimed for primary rather than secondary outcomes	4
Sample characteristics	**Sample size per subgroup:** Credibility increases with sample size per subgroup and balance of sample size across subgroups	13
	Prognostic balance within subgroups: Credibility is higher if, within each subgroup, prognostic factors are balanced	7
	Measurement error: Credibility is higher if the effect modifier is measured without error, e.g. no misclassification	12
Analysis characteristics	**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
	Multiplicity addressed: Credibility is higher if investigators accounted formally or informally for multiplicity *(data-dredging, data-mining)*	45
	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
	Functional form considered: [c] Credibility is higher if the researchers considered the functional form of the model for continuous effect modifiers, e.g. linear or logarithmic relationships	12
	Categorization: [c] Credibility is higher if continuous effect modifiers are not categorized but analyzed as a continuum	10
	Justified cut point: c If continuous effect modifiers are categorized, credibility is higher if the researchers justify the threshold, ideally a priori	1
	Random effects model: b Credibility is higher if the analysis accounts for true variation between studies (within subgroups) by applying a random effects model (Hierarchical model, multilevel model)	9
	Scale dependence: Credibility is lower if effect modification depends on the scale / effect measure	7

Table 2 (continued)

	Sensitivity analysis: Credibility is higher if a sensitivity analysis suggests robustness to relevant assumptions such as cut points or type of model	5
	Bayesian analysis: Credibility is higher if priors were explicitly specified and incorporated using Bayesian methods *(vs. informal account of prior knowledge)*	3
	Shrinkage applied: Credibility is higher if investigators applied shrinkage methods (weighted average of overall effect and subgroup specific effect)	2
Numerical results	**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
	Large effect: Credibility is higher if the effect modification is large	6
	Dose-response: [c] Credibility is higher if there is a dose-response relationship across ordered levels of an effect modifier	2
Contextual considerations	**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
	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
	Prior probability: Credibility increases with the prior probability of the effect modification being true *(prior knowledge, strength of hypothesis)*	16
	Pre-specified 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
	Expert input: Credibility is higher if content expert were involved in the selection of candidate effect modifiers	5
	Consistent across studies: Credibility is higher if the effect modification is consistent across independent studies *(reproducibility, replicability)*	30
	Indirect evidence: Credibility is higher if indirect evidence supports the effect modifier, e.g. evidence from animal studies, observational studies, related populations, or related interventions *(as opposed to theory only or replication in same type of study)*	20
	Consistent across outcomes: Credibility is higher if the effect modification is consistent across related outcomes	7
	Overall effect is significant: Credibility is higher if the overall effect is statistically significant *("positive" vs "negative" trial)*	14
	Overall effect is at low risk of bias: Credibility is lower if the overall treatment effect is at low risk of bias	7
Transparency	**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 pre-specification but without further specification (pre-specification of what? Could refer to the effect modifier, the outcome, the cut point, the model, the test, or test results)

[b] Applies to meta-analysis only

[c] Applies to continuous or ordinal effect modifiers only

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[16]	users and investigators[b]	20	MA	yes, no	no
Burke 2015[40]	n.s.	3	RCT	n.s.	no
European Medicines Agency 2014[39]	users	4-5 [f]	RCT	yes / no [e,f]	no
Koch 2014[41]	n.s.	7	RCT	n.s.	no
Wang 2014a[42]	users	11	RCT	implicitly yes / no	no
Desai 2014[43]	investigators	17	RCT	n.s.	no
Sun 2014[44]	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[45]	investigators	7	RCT	n.s.	no
Pincus 2011[2]	users	5	RCT	implicitly yes / no	systematic survey and consensus study
Cochrane Handbook 2011[46]	n.s.	5	MA	implicitly yes / no	no
Sun 2010[47]	users	4	RCT and MA	implicitly yes / no	no
Kent 2010a[48]	investigators	5	RCT	n.s.	no
Fernandez 2010[21]	users and investigators [g]	14 [g]	RCT	n.s.	no
Dijkman 2009[49]	users	16	RCT	implicitly yes / no	no
Wang 2007[23]	investigators [h]	6	RCT	n.s.	no
Fletcher 2007[50]	users	3	RCT	implicitly yes / no	no
Aulakh 2007[25]	users	6	RCT	n.s.	no
Koopman 2007[24]	investigators	6	individual patient 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[51]	n.s.	11	RCT	n.s.	no
Cook 2004[52]	users	12	RCT	implicitly yes / no	no
Moreira 2002[53]	users	8	RCT	n.s.	no
Brookes 2001[9]	users and investigators [i]	15 [i]	RCT	n.s.	no
Oxman 1992[54]	users	7	RCT and MA	implicitly yes / no	no
Yusuf 1991[55]	investigators	15	RCT	n.s.	no

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

[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 judgement

[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.

[f] 7 items for users, 5 for investigators, 2 for editors

[h] Reporting guideline

[i] 11 items for investigators, 4 for users

Appendix A: Search strategies

Search for journal publications in MEDLINE and Embase (Ovid):

1 (subgroup* or "sub group" or "sub groups").ti.
2 (subset* or "sub set" or "sub sets").ti.
3 (effect modif* or heterogen* or interaction* or moderator* or stratif* or strata or stratum).ti.
4 1 or 2 or 3
5 clinical trials as topic.mp. or exp Clinical Trials as Topic/
6 meta-analysis as topic.mp. or exp meta-analysis/
7 (metaanaly* or meta-analy* or meta analy* or meta-regression or “meta regression”).ti.
8 5 or 6 or 7
9 4 and 8
10 exp Drug Interactions/
11 9 not 10

Search for Textbook chapters in WorldCat.org:
((kw:"subgroup analys#s") OR (kw:"effect modif*")) OR ((kw:interaction OR kw:”heterogeneity") AND (kw:meta-analys#s OR kw:randomi#ed OR ti:epidemiology))
Then activate filters “eBook” and “Print book”
Explanation: “kw” stands for key words and includes search of chapter titles; “ti” applies the search term to book titles only.

Appendix B: Reported credibility criteria, associated rationales, caveats, simulation studies, and references

	Criterion *(related terms)*	**Reported rationales and caveats**
Design Characteristics	**Pre-specification of analytic details (n=39):** Credibility is higher if analytic details such as cut points, time-points, or statistical methods have been pre-specified prior to the analysis.[1,3,5,7,9,21,23,41-43,46,48,49,51,53,55-78]	1) Pre-specifying analytic details can help limits the number of analyses and thus multiplicity problems (as long as the analysis plan was specific, the number of analyses small, and the investigators adhered to the plan)[9,21,23,48,57,59,65,73] 2) Pre-specification of analytic details can decrease the risk of selective reporting (as long as the analysis plan is available and the investigators report all analysis accordingly)[5] 3) Ability to pre-specify an analysis in detail suggests prior knowledge (see "a priori hypothesis")[60,79] *Caveat: statements of pre-specification can often not be verified in study protocols[19]* *Caveat: Pre-specification of analytic details without knowledge of the data may not be plausible[57]*
	Small number of candidate effect modifiers (n=38): Credibility is higher if only a small number of effect modifiers have been tested. [1,5-7,9,16,20,27,28,30,31,40,41,43,44,46,47,49,51,53-55,57,59,69,74,75,80-89]	1) Avoids multiplicity issues by design[6,43,44,55,81,88] 2) Preserves power (see "power")[40,51,55] 3) As investigators add subgroup analyses with less and less evidence or theory to support them, the average prior probability for an effect modification falls, thus further reducing credibility.[40,82]
	Within versus between study comparison (n=25): Credibility is higher if inferences regarding effect modification are based on within-study analyses rather than a comparison of main effects across studies**.** *(individual patient data vs aggregate data meta-analysis)* [3,7,22,44,46,47,54,57,64,68,72,80,90-101]	1) Within-study comparison avoids aggregation bias[7,64,94,95,97] 2) Within-study comparison has usually more power because of a larger sample size and a wider spectrum of values[57,91,92,99] 3) Inferences based on between-study comparisons are particularly susceptible to confounding.[44,57,64,68,93,98,102] 4) Comparison of within-study analyses facilitates a formal assessment of consistency across studies (see "consistency across studies")[102] Simulations studies: [92,97]
	Baseline characteristic vs. characteristic measured after randomization (n=24): 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. [2,5,6,20,39,43,44,47,49,52-55,59,60,66,67,77,103-108]	If subgroup membership is influenced by an effect of the intervention, this biases the prognostic balance between intervention and control within subgroups. [6,43,44,47,52,54,55,59,66,67,104-106]
	Power (n=23): Credibility increases with the prospective power to detect the effect modifier [3,5,21,26,31,36,40,42,49,51,57,59,85,92,109-115]	More power increases the likelihood of a claim being true. The analogy of a cross-table for a diagnostic test provides an explanation, with a positive test result being a significant test of interaction: sensitivity corresponds to the prospective power (i.e. the ability of the test to detect a true interaction, e.g. 50%) and specificity corresponds to 1 - α (i.e. the ability of the test to correctly dismiss a spurious interaction due to randomness, e.g. 95%). Our interest is in the probability of a significant test being true which corresponds to the positive predictive value. The latter increases with sensitivity/power.[40,110] (The issue of low power is widely acknowledged in the context of false-negative conclusions, less so in the context of false positive) Simulation studies: [89,116]

Appendix B (continued)

	Regression on control group risk (n=6): Credibility is lower if, in a meta-analysis, the effect modifier is defined by the risk of an outcome in the control group. [57,69,93,102,121,122]	A number of papers have shown that regression of the effect on the control group risk in a meta-analysis introduces a bias (regression to the mean). Methods are available to handle the problem.[102,104,108,121,122]
	Primary outcome (n=4): Credibility is higher for effect modifiers claimed for primary rather than secondary outcomes[23,31,74,83]	1) Trials have usually most power for the primary outcome (see "power") [74] 2) Confining analyses of effect modification to the primary outcome limits the number of analyses (see "small number of candidate effect modifiers")[83]
Sample characteristics	**Sample size per subgroup (n=13):** Credibility increases with sample size per subgroup and balance of sample size across subgroups [16,36,40-42,46,49,51,75,115,123-126]	This maximizes power (see "power")[40,51] Simulation studies: [9,89,110,125]
	Prognostic balance within subgroups (n=7): Credibility is higher if, within each subgroup, prognostic factors are balanced[1,39,49,51,70,73,119]	As for the overall trial, randomization is assumed to balance the prognosis between arms within a subgroup. If randomization fails, e.g. in small subgroups, this may result in spurious inferences (related to "stratified randomization").[49,73]
	Measurement error (n=12): Credibility is higher if the effect modifier is measured without error, e.g. no misclassification. [2,6,7,42,55,71,78,89,95,127-129]	Measurement error can lead to biased conclusions. Random misclassification would most likely dilute an effect modification, which is most relevant in the context of false-negative.[2,6,78,95,127,128]
Analysis characteristics	**Interaction test (n=54):** 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)*[1,2,5-9,16,20,21,23,25-28,30,31,39,42-50,52-56,59,61,65,66,69,70,73,74,80,83-85,89,106,107,113-115,124,130-132]	The interaction test directly addresses random error.[9,30,54,59,124] Simulation studies: [9,132]
	Multiplicity addressed (n=45): Credibility is higher if investigators accounted formally or informally for multiplicity (data-dredging, data-mining)[1,5,8,9,21,23,25,28,41-43,45,48,49,51,52,55,60,61,65-67,69,73,74,80,83,85,88,89,98,106,107,109-111,113,119,124,128,129,133-139]	Multiple analyses increase the risk for identifying random results. There are a number of methods available to address multiplicity such as reducing the number of analyses, considering lower thresholds for significance, multivariable analysis, or using composite variables.[23,45,49,55,61,67,69,73,85,88,98,106,110,128,133,134,136] Caveat: Adjustment for multiplicity reduces power[85,129,135] Simulation studies: [110,140]
	Effect modification persists after adjusting for other effect modifiers (n=17): 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)[1,3,7,8,20,46,47,57,68,80,88,128,135,141-144]	1) Statistical independence makes confounding less likely. Credibility is higher if the effect modifier is a cause of the outcome as opposed to being a proxy for the cause (e.g. age as a proxy for comorbidity). Treatment decisions may be unreliable if the proxy is unreliable (e.g. if age is a poor proxy for comorbidity).[38,68,80,85,93,135,144,145] 2) multivariable analyses control the type 1 error (see "multiplicity addressed")[85] Caveat: Some authors argue that a causal interpretation is not necessary as the aim is identification of target subgroups[88,146], but others disagree.[38] Simulation studies: [38]
	Functional form considered (n=12): Credibility is higher if the researchers considered the functional form of the model for continuous effect modification, such as linear or logarithmic relationships. [39,51,57,71,128,134,147-151]	1) Model misspecification is a potential source of bias as the apparent interactions may be driven by a few influential observations.[148,151] 2) Model misspecification may cause a loss of power.[134] Caveat: model selection adds another layer of multiplicity (REF) Simulation studies: [37,148]
	Categorization (n=10): Credibility is higher if continuous effect modifiers were not categorized (vs. continuous)[4,16,40,51,53,71,93,98,134,135,149]	1) Arbitrary selection of cut-points is associated with multiplicity issues.[98,134,135] 2) Loss of power (see "power")[40,93,98,134]

Appendix B (continued)

	Justified cut point (n=1): If a continuous effect modifier is categorized, credibility is higher if the researchers justify the threshold, ideally a priori[60]	not reported
	Random effects model (n=9): Credibility is higher if the analysis accounts for true variation between studies (within subgroups). *(Hierarchical model, multilevel model)*[46,57,68,95,101,110,125,129,152]	Assuming true variation between individuals (or studies) within subgroups is usually plausible and has implications for both summary effects and standard error. In most instances, significant results are harder to achieve with a random effects model, and the approach would protect against over-interpretation.[57,68,95,110,152] *Caveat: estimation of within group dispersion within a subgroup may be imprecise if the number of studies is small.*[68] Simulation studies: [101,110,125]
	Scale dependence (n=7): Credibility is lower if effect modification depends on the scale.[39,49,59,71,76,143,147]	Effect modification is usually scale dependent. Some authors consider effect modifiers that disappears by modifying the scale as artificial.[49,143,147] Caveat: Testing effect modification on several scales creates a multiplicity problem[71] Simulation studies: [71,85]
	Sensitivity analysis (n=5): Credibility is higher if a sensitivity analysis suggests robustness to relevant assumptions such as cut points or type of model[39,43,45,128,129]	Sensitive to assumptions means more likely spurious[43,128,129]
	Bayesian analysis (n=3): Credibility is higher if priors were explicitly specified and incorporated using Bayesian methods (vs. informal account of prior knowledge)[40,65,69,153]	1) Priors allow explicit quantification of prior knowledge/uncertainty[69] 2) Automatic adjustment for multiple comparison[65] Caveat: prior distributions may not reasonably represent the prior beliefs[69] Simulation studies: [154]
	Shrinkage applied (n=2): Credibility is higher if investigators applied shrinkage methods (weighted average of overall effect and subgroup specific effect)[62,129]	1) Tempers optimism regarding the size of effect modification [124,135] 2) A method to control for multiplicity issues[129,135] Simulation studies: [124]
Numerical results	**Quantitative vs qualitative (n=12): Credibility is higher if the effect modification is quantitative (direction of effect consistent across levels of the effect modifier) rather than qualitative (direction varies by levels of the effect modifier).**[28,30,41,44,55,56,58,59,80,106,147,155]	1) Qualitative interactions have a low prior probability because they are generally rare[30,155] 2) Qualitative interactions would require a rather complex causal rationale for justifying both benefit for one subgroup but harm for another subgroup[106] Caveat: There are specific situations where qualitative interactions are plausible, e.g. in trials investigating targeted anti-cancer drugs[156]
	Large effect (n=6): Credibility is higher if the effect modification is large.[3,39,47,54,75,133]	Not reported
	Dose-response (n=2): Credibility is higher if there is a dose-response relationship across ordered levels of an effect modifier[52,106]	Not reported
Contextual considerations	**A priori hypothesis (n=49): Credibility is higher if the investigators stated a hypothesis prior to performing the study, lower if an explanation arose after data analysis. (vs. post-hoc explanation; confirmatory vs exploratory; hypothesis testing vs hypothesis generating)**[2,5-7,9,20,21,25-27,39-50,52,54,55,58,63,64,67,72,74-76,78,90,93,107,109,111,112,114,128,129,135,139,153,154,157,158]	1) Being able to specify a hypothesis a priori makes prior knowledge or evidence more likely; credibility increases with prior probability.[21,40,55] 2) When the explanation arose post hoc, investigators likely considered many possible explanations. This creates a multiplicity (and possibly selective reporting) problem.[5,7,49,50,54,55,90]

Chapter 4: Development of a new Instrument to assess the Credibility of Effect Modification ANalyses (ICEMAN) in randomized controlled trials and meta-analyses

We are currently formatting this manuscript for submission to BMJ.

Stefan Schandelmaier, *methodologist*[1,2]; Matthias Briel, *associate professor* [1,2]; Ravi Varadhan, *associate professor* of Oncology Biostatistics[4]; Christopher H Schmid, *professor of biostatistics*[10]; Niveditha Devasenapathy, *associate professor*[11]; Rodney A Hayward, *professor*[12]; Joel Gagnier, *associate professor*[13]; Michael Borenstein, *statistician*; Geert JMG van der Heijden, *clinical epidemiologist*[8]; Issa J Dahabreh, *assistant professor*; Xin Sun, *professor*[15]; Willi Sauerbrei, *professor in medical biometry*[16]; Michael Walsh, *associate professor*[1,3,7]; John PA Ioannidis, *professor*[9]; Lehana Thabane, *professor*[1,5,6]; Gordon H Guyatt, *distinguished professor*[1,3]

1) Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada
2) Institute for Clinical Epidemiology and Biostatistics, Department of Clinical Research, Basel University, Basel, Switzerland
3) Department of Medicine, McMaster University, Hamilton, Ontario, Canada
4) Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205
5) Departments of Pediatrics and Anesthesia, McMaster University, Hamilton, Ontario, Canada
6) Biostatistics Unit, St Joseph's Healthcare—Hamilton, Ontario, Canada
7) Population Health Research Institute, Hamilton Health Sciences/McMaster University, Hamilton, Canada
8) Department of Social Dentistry, Academic Center for Dentistry Amsterdam, University of Amsterdam and VU University Amsterdam
9) Departments of Medicine, of Health Research and Policy, of Biomedical Data Science, and of Statistics and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, CA, USA
10) Center for Evidence Based Medicine, School of Public Health, Brown University, Providence, Rhode Island, USA
11) Indian institute of Public Health-Delhi, Public Health Foundation of India, New Delhi, India
12) Department of Internal Medicine and Department of Health Management and Policy, University of Michigan, Ann Arbor, Michigan, USA
13) Department of surgery and epidemiology, University of Michcigan, Ann Arbor, Michigan, USA
14) Department of Health Services, Policy & Practice, and Department of Epidemiology, Center for Evidence Synthesis in Health, School of Public Health, Brown University, Providence, Rhode Island, USA
15) Chinese evidence-based medicine centre, Sichuan University, China
16) Institute for Medical Biometry und Statistics, Albert-Ludwigs-University Freiburg, Germany

Abstract

Background: Some randomized controlled trials (RCTs) and meta-analyses make claims of *effect modification* (synonyms: *subgroup effect* or *interaction*) in which the effect of an intervention varies by another variable (the *effect modifier*). Deciding on the credibility of an apparent effect modification presents challenges.

Objective: To develop a formal instrument for assessing the credibility of effect modification analyses (ICEMAN) in an RCT or meta-analysis of RCTs.

Methods: Following a stepwise process, we developed a detailed concept; identified candidate credibility considerations in a systematic literature survey; together with leading 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 (five for RCTs, eight for meta-analyses) with four response options, one optional item for additional considerations, and a credibility rating on a visual analogue scale ranging from very low to high credibility. An accompanying manual provides rationale, detailed instructions, and examples from the literature. Seventeen potential users tested the instrument. Implementing their suggestions improved the user-friendliness of the new instrument.

Conclusions: ICEMAN is a rigorously developed instrument to rate the credibility of apparent effect modification. The content and presentation has been optimized for trial investigators, systematic reviewers, journal editors and others who are interpreting claims of effect modification.

Box 1: Variation in terminology

Synonyms for effect modification	Synonyms for effect modifier
• Subgroup effect • Interaction • Moderation • Differential effect • Heterogeneity of treatment effects	• Subgrouping variable • Predictor of treatment response • Moderator

Note that some methodologists conceptually distinguish effect modification from interaction,[1,2] but most authors use the terms synonymously referring to effect modification.

Introduction

Investigators of randomized clinical 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, type of intervention or risk of bias.[3-16] The terminology varies; Box 1 presents the alternatives currently in use.

Not infrequently, investigators claim to have identified an effect modification. Meta-research suggests that 14-26% of RCTs and meta-analyses emphasize a potential effect modification in their abstract or discussion.[6-11,13]

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

At the same time, however, mistaken claims of effect modification may compromise optimal patient care. Examples of putative effect modification refuted by subsequent evidence include claims that aspirin is effective for secondary stroke prevention in men but not women; [17-19] that antihypertensive treatment reduced stroke and heart failure in younger patients but not in the elderly;[20,21] or that ticlopidine is superior to aspirin in preventing cardiovascular events in blacks but not in whites.[22,23] A substantial number of examples of effect modification subsequently proved spurious.[24-26] Applying an effect modification that is spurious may either cause harms through administration of ineffective treatment or deny beneficial therapies to patients, and will likely increase health care costs.

Numerous theoretical analyses and simulation studies explain why claims of effect modification are often misleading.[27] The fundamental reason is chance: if an effect truly is the same in all subgroups of patients, testing a sufficient number of candidate effect modifiers will inevitably reveal an apparent, but misleading, effect modification. Other reasons that contribute to

spurious claims include selective reporting,[7,9] lack of hypotheses and supporting evidence,[7,9,28], and failure to use a proper test of interaction.[7,10] In the context of meta-analysis, uncertainty may be further increased because of confounding on the study-level.[29-31]

Nevertheless, some claims of effect modification – probably a small minority[3] – will represent true effects. Because most claims arising for single studies 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 versus real claims.

In response, methodologists first suggested credibility criteria in the early 1990s.[32,33] Since then, a total of 30 groups have suggested sets from 3 to 21 criteria.[27] Aside from variability certain to sow confusion, previous sets suffer from suboptimal presentation, resulting in ambiguity in applying criteria. Some criteria – for instance whether the effect modifier was one of a small number tested[34] – involve substantial subjectivity, and would therefore benefit from more detailed guidance. Most importantly, none of the previous sets of criteria followed a rigorous development process or underwent serious user-testing before publication.[27]

We therefore systematically developed ICEMAN to provide a methodologically rigorous, user-friendly instrument to assess the credibility of apparent effect modification.

Development process

ICEMAN development consisted of four steps recommended for the development of measurement instruments in general[35] and specifically for research quality appraisal tools[36]: A) clarifying the concept, B) systematic literature survey to identify existing instrument and candidate credibility criteria, C) consensus study among experts to select criteria and develop items, and D) formal user-testing.

A) Concept
Members of the steering committee (StS, GG, MB, XS, MW, LT) developed the concept of the instrument based on their methodological expertise and practical experience in performing and interpreting analysis of effect modification, and development of quality appraisal tools. The draft concept specified the following:

Aim of the new instrument: To assist users in assessing the credibility of a claimed effect modification (rather than claims that an effect modification is absent, which would require different criteria).
Definition of credible effect modification: Effect modification means that the effect of an intervention on an outcome varies by levels of another variable. An effect modification is credible if it is very likely true, i.e. not the result of chance or bias. We also clarified that patient-importance is not part of the credibility assessment, because considerations of

importance depend on context, and that credibility can be assessed on any scale, e.g. risk ratio or risk difference scales.
Target users: Health care providers, trial investigators, systematic review authors, journal editors, guideline developers, and health policy makers.
Type of studies: The instrument will address RCTs and meta-analyses of RCTs.
Format: The core instrument will consist of signaling questions with response options, and no more than 8-12 items.
Responsiveness: The instrument should be responsive, i.e. studies should vary in the extent to which they meet criteria. Overly strict or lenient items that do not vary are useless for distinguishing more from less credible effect modification.
Overall credibility: The instrument should conclude with a summary rating that expresses the overall credibility of the proposed effect modification on a continuum ranging from very low to high credibility.

This concept influenced all steps in the development process including the design of the systematic survey, the selection of experts (we presented the concept in our invitation), the selection and development of candidate items, and the design of the user-testing. We refined the concept throughout the project. Major developments were the decisions to make two separate versions for RCTs and meta-analyses of RCTs and to include an optional item addressing additional credibility considerations.

B) Systematic survey of the methodological literature: The objectives of the systematic survey, presented in a separate publication,[27] were to identify 1) existing instruments for assessing credibility of effect modification and verify that none of them satisfied our concept, 2) candidate credibility criteria, defined as characteristics of an analysis of effect modification suggested to either in- or decrease credibility, and 3) leading experts in the field, defined as first, second, or last authors of two or more eligible publications. The systematic survey included a comprehensive search of journal articles and textbooks; teams of reviewers extracted data in duplicate and performed a formal qualitative synthesis process.[27]

The systematic survey identified 150 eligible publications from which we extracted 36 candidate criteria, 30 previous sets of criteria (none sufficiently reflected our concept) and 40 experts. The survey highlighted which criteria are most popular, most controversial, and more or less supported by a rationale or simulation studies.[27]

C) Expert consensus. The aim of the consensus study – informed by the results of the systematic survey and later the user-testing study – was to identify key criteria for the credibility of effect modification and develop the criteria into a user-friendly and responsive instrument**.**

A colleague not involved in the project and blinded to names randomized the order of the 40 candidate experts. The steering committee invited the first 18 experts, of whom 11 agreed to participate, 4 declined, and 3 did not respond. Of the 11 who initially agreed, 9 participated in the final consensus study; 1 withdrew before the first telephone conference due to over-

commitment, and 1 due to a research focus on observational studies. The final group included 15 members (the 6 members from the core group and 9 external experts). The consensus study took place between March 2018 and February 2019 and consisted of the following steps:

1) Selection of key criteria: StS created a list of the 36 candidate criteria identified in the systematic survey, including for each criterion frequency of reporting and rationale (Appendix 1). Members of the group (excluding StS) reviewed the candidate criteria and rated the importance of each criterion using a 7-point scale with 1 indicating not important at all and 7 indicating highly important for credibility assessment. In addition, group members suggested to the group to merge, drop, or add new criteria. StS summarized the results that then provided the basis for the first video-conference.

2) During the first video-conference (1.5 hours, 11 participants) the group agreed on a concept and decided on criteria that should be definitely included or excluded. The group identified 20 criteria that should be definitely included (some of which we later combined), 8 definitely excluded, and 8 optional (Appendix 1). After the conference, all group members received a detailed summary of the discussion and decisions and had the opportunity to provide additional written feedback to all group members.

3) Based on the initial criteria selection, the core group developed a first draft of the instrument and transformed the credibility considerations into explicit items composed of signaling questions (Appendix 1), each item with four response options, and illustrative examples for each response option. Initially, we planned to create a single instrument applicable to individual RCTs, aggregate data meta-analyses, and individual participant data meta-analyses. Because a single version proved excessively complex, the group decided to create two separate versions, one for individual RCTs (6 initial core items) and one for meta-analyses (9 initial core items). In addition, the group offered a set of preliminary considerations, included less important credibility criteria as a list of optional considerations, and drafted a final item to assess overall credibility using a visual analogue scale. Where possible, we used a format similar to other popular quality appraisal instruments such as the Cochrane risk of bias tool[37] or the GRADE evidence to decisions frameworks.[38] In preparation for the second video conference, all group members had the opportunity to comment on the draft.

4) The aim of the second video conference (1.5 hours, 11 participants in two group sessions and 3 individual discussions with experts who could not attend group sessions) was to find a consensus on a general structure of the instrument, including preliminary considerations, core items, optional considerations, format of the overall rating, but not yet precise wording. Main discussion points included approaches to make response options explicit, e.g. whether to use p-values, thresholds for p-values, relevant sources of multiplicity, issues of threshold selection for continuous effect modifiers, how to frame optional considerations, and whether individual patient-data meta-analyses should be combined with the version for RCTs or meta-analyses. The group agreed on the main structure, number of response options, the design of the overall rating item, and several statistical details (Appendix 1). After the conference, all group

members received a summary of the decisions and had the opportunity to provide additional written feedback to all group members.

5) The last part of the consensus study included the following: 1) We created a detailed manual and sought, for each response option, a supporting example of an apparent effect modification published in the medical literature. Applying drafts of the instrument to potential examples led to a number of improvements. 2) We presented the instrument at the annual Cochrane Conference in Edinburgh 2018; a discussion with attending methodologists led to a refined concept on which we elaborate in the manual. 3) Most relevant suggestions for improvement came from the user testing (see next section).

Through the last phases of development, we periodically circulated updated versions to the experts, inviting them to provide comments.

Appendix 1 documents major developments throughout the consensus study.

D) User-testing

The aim of the user-testing was to identify the challenges experienced by members of the target audience in applying an advanced draft of ICEMAN to a published claim of an effect modification that we provided. Each user received the abstract and full text of an RCT or meta-analysis in which authors claimed one effect modification, the appropriate version of ICEMAN, and the manual. To ensure variation across the range of possible claims, we selected claims that we judged to have very low (n=4), low (n=5), moderate (n=4) or high (n=4) credibility of effect modification. We included 10 RCTs and 7 meta-analyses.

We recruited 17 potential users from three main sources: corresponding authors of randomly selected Cochrane reviews published after July 2017 (7 participants); corresponding authors of randomly selected RCTs published in Lancet, JAMA, BMJ, Annals of Internal Medicine, or PLOS Medicine after July 2017 (3 participants); and 5 journal editors and 2 trial investigators from personal networks who were not involved in instrument development and not located at McMaster University. We continued to enroll users until they did not identify any new major limitations. The users varied with respect to gender (which we used as a stratification factor), background, and familiarity with issues of effect modification (Table 1).

One of two investigators (StS or ND) interviewed users immediately after they had applied ICEMAN following a semi-structured interview guide. The guide included pre-defined open questions (e.g. “What was your experience when you applied the first item?”), but allowed interviewer or interviewee to extend on relevant ideas that came up during the interview (Appendix 2). Interviews lasted from 25 to 70 minutes (median 37 minutes). The interviewers immediately transcribed the recorded interviews and extracted positive comments, negative comments, and suggestions for improvement using qualitative data analysis software www.dedoose.com.

Appendix 2 lists the number of positive and negative comments and major changes resulting from the comments. Critical comments that we could not address included that some users may have to seek assistance from a statistician when using ICEMAN; ICEMAN does not address uncertainty arising from potential conflicts of interest of authors, e.g. when subgroup analyses of otherwise uninteresting results are motivated by the desire to publish secondary papers; or that ICEMAN does not fit one page. Frequent positive comments included that users would be happy to use the ICEMAN again and found it instructive, useful, and easy to follow.

We updated the instrument three times during the ongoing user-testing (after 7 interviews, 12 interviews, and 15 interviews) before the consensus group finalized the instrument and manual.

The final instrument

Appended are the final versions of ICEMAN for an individual RCT (Appendix 3), a meta-analysis of RCTs (Appendix 4), and the manual for both versions (Appendix 5).

Both versions start with a set of five preliminary considerations designed to link ICEMAN to a study and, if available, a study protocol; define the effect modification under consideration through a single outcome, effect measure, and effect modifier; and alert users that ICEMAN may not apply to effect modifiers measured after randomization (see manual for more details).

The version for meta-analysis includes 8, the version for individual RCTs 5 core signaling questions, 4 of which overlap among the two versions (Table 2).

For each response option, ICEMAN provides four response options. The response options differ in wording but share the same order and logic such as definitely no, probably no or unclear, probably yes, and definitely yes. Response options on the left indicate definitely or probably reduced credibility, response options on the right probably or definitely increased credibility. We included response option probably no with unclear to cover situations with insufficient information (Appendices 3 and 4).

After the core questions, one optional question allows additional credibility considerations such as results from a sensitivity analysis, a dose-response relationship, or other considerations that are difficult to ascertain, are less relevant, or seldom apply. Additional consideration are optional, and can reduce or increase credibility (Appendices 3 and 4).

ICEMAN concludes with an overall credibility rating presented on a visual analogue scale (VAS) on which users place a mark. The VAS is divided into four areas labelled as very low, low, moderate, and high credibility. The areas roughly correspond to probabilities of <25%, 25-50%, 50-75%, and >75% that the effect modification truly exists. To aid interpretation, the final item provides suggestions – rather than an algorithm – for deriving overall credibility from the responses to the previous questions (Appendices 3 and 4).

The manual provides, for each element of ICEMAN, more detailed explanations, a rationale with key references and examples for support, and examples of completed credibility assessments. In addition, the manual includes practical suggestions for presentation and using ICEMAN in combination with other risk of bias instruments and GRADE. A final chapter elaborates on conceptual considerations (Appendix 5).

Discussion

ICEMAN provides a systematically developed instrument to assess the credibility of effect modification proposed in an RCT or meta-analysis of RCTs, providing a guide for users to consider the implications of proposed effect modifications for patient care. ICEMAN was developed by experts in assessment of effect modification building on a systematic survey of the entire relevant literature; provides versions for randomized trials and meta-analyses; is succinct (5 core items for the RCT version and 8 for the systematic review version); is structured (preliminary considerations, signaling questions, overall rating); and provides a detailed instruction manual to guide its use (Appendices 3-5).

A possible limitation of ICEMAN is that, to optimize the reliability of application, formulating four response options required specification of threshold values for credibility with respect to the number or studies in a meta-analysis, p-values, and number of candidate effect modifiers. These thresholds suffer from some degree of arbitrariness, and experts initially disagreed on the thresholds. Particularly controversial within our group were thresholds for interaction p-values, although the group finally found compromises acceptable to all. Perhaps reassuring, none of the participants of the user-testing study mentioned concerns with chosen thresholds, and those who made a comment appreciated the explicit thresholds. Nevertheless, some users might disagree with the chosen thresholds.

Another potential limitation is that the core questions do not include all credibility considerations that experienced analysts might deem relevant, in particular for complex analyses such as modeling of continuous effect modifiers[39,40] or data-driven algorithms for subgroup discovery.[41,42] For instance, experienced analysts might question the appropriateness of statistical models underlying tests for interaction,[28,43,44] differ in their conceptualization of family wise error,[45] or may want to consider the correlation structure between multiple effect modifiers.[46] Even for such users, ICEMAN will provide a useful starting point for analysis of effect modification. For instance, if the core questions suggest low or very low credibility, it is very unlikely that investing in more complex analyses could substantially increase credibility; if the core questions suggest moderate credibility, ICEMAN provides an optional item in which users can refer to additional considerations.

Some properties of ICEMAN remain uncertain. In a future project, we will 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, if there will ever be sufficient data available to investigate validity if we consider independent replication

the reference standard for establishing a true effect modification. A recent analysis has shown that attempts to replicate effect modification claimed in RCTs are extremely rare.[26] Therefore, we invite users of ICEMAN 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-point, potentially assess the extent to which the claims withstood or failed tests of replication. This will also allow better calibration of the different categories of assigned overall credibility. In addition, we will continue to evaluate the performance of ICEMAN in practice. We invite users to report difficulties or suggestions for improvement for consideration in future modifications of ICEMAN (please write to corresponding author).

In summary, ICEMAN provides a rigorously developed and thoroughly user-tested instrument for judging the credibility of putative effect modification in RCTs and meta-analyses. We anticipate that both authors and target audiences of RCTs and meta-analyses, and other groups including journal editors, will find the structured assessment of credibility of proposed effect modification helpful.

Acknowledgements: We thank the 17 users who tested drafts of ICEMAN and the manual and shared their valuable experience with us in detailed interviews. We also thank the authors of the many well written methodological publications about effect modification that provided a comprehensive knowledge base. In addition, we thank Noel Weiss, Ian Shrier, and X referees for their constructive comments.

Funding: StS personal grants: The Swiss national science foundation P300PB_16475; Gottfried and Julia Bangerter foundation; Freiwillige akademische Gesellschaft Basel; Forschungsfonds Basel University

References

1.VanderWeele TJ. Explanation in causal inference. Methods for mediation and interaction. 1 ed. New York: Oxford University Press; 2015.
2.Corraini P, Olsen M, Pedersen L, Dekkers OM, Vandenbroucke JP. **Effect modification, interaction and mediation: an overview of theoretical insights for clinical investigators.** Clin Epidemiol. 2017;9:331-8.
3.Schuit E, Li AH, Ioannidis JPA. **How often can meta-analyses of individual-level data individualize treatment? A meta-epidemiologic study.** International journal of epidemiology. 2018.
4.Gabler NB, Duan N, Raneses E, Suttner L, Ciarametaro M, Cooney E, et al. **No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals.** Trials. 2016;17(1):320.
5.Simmonds M, Stewart G, Stewart L. **A decade of individual participant data meta-analyses: A review of current practice.** Contemporary clinical trials. 2015;45(Pt A):76-83.
6.Zhang S, Liang F, Li W, Hu X. **Subgroup Analyses in Reporting of Phase III Clinical Trials in Solid Tumors.** Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2015;33(15):1697-702.
7.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(6):e0128804.
8.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.** European journal of cancer. 2015;51(18):2732-9.
9.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.
10.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.
11.Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. **Statistics in medicine--reporting of subgroup analyses in clinical trials.** The New England journal of medicine. 2007;357(21):2189-94.
12.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.** Journal of clinical epidemiology. 2007;60(10):1002-9.
13.Hernandez AV, Boersma E, Murray GD, Habbema JD, Steyerberg EW. **Subgroup analyses in therapeutic cardiovascular clinical trials: are most of them misleading?** American heart journal. 2006;151(2):257-64.
14.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.** Clinical orthopaedics and related research. 2006;447:247-51.
15.Moreira ED, Jr.; Stein, Z.; Susser, E. **Reporting on methods of subgroup analysis in clinical trials: a survey of four scientific journals.** Brazilian journal of medical and biological research = Revista brasileira de pesquisas medicas e biologicas / Sociedade Brasileira de Biofisica [et al]. 2001;34(11):1441-6.

16.Assmann SF, Pocock SJ, Enos LE, Kasten LE. **Subgroup analysis and other (mis)uses of baseline data in clinical trials.** Lancet. 2000;355(9209):1064-9.
17.Fields WS, Lemak NA, Frankowski RF, Hardy RJ. **Controlled trial of aspirin in cerebral ischemia.** Stroke; a journal of cerebral circulation. 1977;8(3):301-14.
18.Canadian Cooperative Study G. **A randomized trial of aspirin and sulfinpyrazone in threatened stroke.** The New England journal of medicine. 1978;299(2):53-9.
19.**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.** Bmj. 1994;308(6921):81-106.
20.Amery A, Birkenhager W, Brixko P, Bulpitt C, Clement D, de Leeuw 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: sub-group analysis on entry stratification.** J Hypertens Suppl. 1986;4(6):S642-7.
21.Gueyffier FB, C.; Boissel, J. P.; Schron, E.; Ekbom, T.; Fagard, R.; Casiglia, E.; Kerlikowske, K.; Coope, J. **Antihypertensive drugs in very old people: a subgroup meta-analysis of randomised controlled trials. INDANA Group.** Lancet. 1999;353(9155):793-6.
22.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(1):27-31.
23.Gorelick PB, Richardson D, Kelly M, Ruland S, Hung E, Harris Y, et al. **Aspirin and ticlopidine for prevention of recurrent stroke in black patients: a randomized trial.** Jama. 2003;289(22):2947-57.
24.Rothwell PM. **Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation.** Lancet. 2005;365(9454):176-86.
25.Guyatt G. Users' guides to the medical literature : a manual for evidence-based clinical practice. New York: McGraw-Hill Education; 2015.
26.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.
27.Schandelmaier S, Chang Y, Devasenapathy N, Devji T, Kwong JSW, Colunga Lozano LE, et al. **A systematic survey of suggested criteria for assessing the credibility of effect modification in randomized controlled trials or meta-analyses.** under review in Journal of Clinical Epidemiology.
28.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.** International journal of epidemiology. 2016;45(6):2184-93.
29.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.** International journal of technology assessment in health care. 2008;24(3):358-61.
30.Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. **Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?** Bmj. 2017;356:j573.
31.Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI, Anti-Lymphocyte Antibody Induction Therapy Study G. **Individual patient- versus group-level data meta-regressions for**

the investigation of treatment effect modifiers: ecological bias rears its ugly head. Statistics in medicine. 2002;21(3):371-87.
32.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(1):93-8.
33.Oxman AD, Guyatt GH. **A consumer's guide to subgroup analyses.** Annals of internal medicine. 1992;116(1):78-84.
34.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.
35.Streiner DL, Norman GR, Cairney J. Health measurement scales : a practical guide to their development and use. Oxford: Oxford University Press; 2015.
36.Whiting P, Wolff R, Mallett S, Simera I, Savovic J. **A proposed framework for developing quality assessment tools.** Systematic reviews. 2017;6(1):204.
37.Higgins J, Sterne J, Savović J, Page M, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie J, Boutron I, Welch V, editors. Cochrane Methods: Cochrane Database of Systematic Reviews 2016;(10 Suppl 1); 2016.
38.Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, 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.
39.Royston P, Sauerbrei W. **Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis.** Statistics in medicine. 2013;32(22):3788-803.
40.Royston P, Sauerbrei W. **Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis.** Statistics in medicine. 2014;33(27):4695-708.
41.Lipkovich I, Dmitrienko A, B. R. D' Agostino S. **Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials.** Statistics in medicine. 2017;36(1):136-96.
42.Lipkovich I, Dmitrienko A, Muysers C, Ratitch B. **Multiplicity issues in exploratory subgroup analysis.** Journal of biopharmaceutical statistics. 2018;28(1):63-81.
43.Royston P, Sauerbrei W. **A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials.** Statistics in medicine. 2004;23(16):2509-25.
44.Borenstein M, Higgins JP. **Meta-analysis and subgroups.** Prevention science : the official journal of the Society for Prevention Research. 2013;14(2):134-43.
45.Dmitrienko A, D'Agostino R, Sr. **Traditional multiplicity adjustment methods in clinical trials.** Statistics in medicine. 2013;32(29):5172-218.
46.Varadhan R, Wang SJ. **Standardization for subgroup analysis in randomized controlled trials.** Journal of biopharmaceutical statistics. 2014;24(1):154-67.

Table 1: Characteristics of participants in the user-testing study

Characteristic	n (total 17)
Female	10
Continent	
US / Canada	7
Europe	6
Asia	3
Australia	1
Current primary professional activity	
Journal editor	5
Health care provider	5
Researcher	5
Statistician	2
Experience in roles (more than one possible)	
Trial investigator	7
Systematic review author	13
Guideline developer	6
Journal editor	8
Highest academic degree	
PhD	6
MSc	6
MD	5
Self-rated familiarity with analyses of effect modification	
Not at all	0
A little	6
Somewhat	7
Very	4

Table 2: Core questions of the two versions of ICEMAN (numbers reflect order of appearance)

Core questions	Meta-analysis of RCTs	Individual RCT
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?	4	1
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?	5	3
Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?	6	4
Did the authors use a random effects model?	7	
If the effect modifier is a continuous variable, were arbitrary cut points avoided?	8	5

Appendix 1: Overview of steps in the consensus study (electronic file)

Initial list of criteria presented to the expert group (number of publications that mentioned this criterion)	Importance ratings provided by 10 members of the group (mean); 7 indicates most important, 1 indicates not important for credibility instrument. Not shown are verbatim comments and suggestions	Decisions made in 1st video conference: item selection and other decisions	First draft of ICEMAN for meta-analyses	First draft if ICEMAN for RCTs	Main discussion points and decisions of the 2nd video conference: focus still on concept rather than wording	Main discussion points in written discussions in remaining development process including written feedback to draft manual
		General decisions: - We will create two separate instruments for individual RCTs and conventional meta-analysis (MA) (not clear yet how IPD MA will fit in) - We will restrict the instrument to effect modifiers measured at baseline - We will include considerations that are not key but sometimes relevant as secondary or additional considerations	The first draft included suggestions for preliminary considerations, core items (9 for meta-analysis, 6 for individual RCT), additional considerations presented as a checklist at the end of the instrument, some framed positively, some framed negatively; and an overall rating designed as a four labelled points connected through a line. Not shown here are the drafts of the four response options with corresponding explanations		General decisions: - We decided to develop two versions, one for individual RCTs, one for meta-analyses of any type - We agreed on the target audience - We agreed to put complex considerations in the manual (which to develop will be the next step) - We agreed on using a VAS to rate the overall. We revised to make the VAS look more continuous and make clear that the four credibility areas are not checkboxes. - We decided the call the instrument ICEMAN	General decisions: - We arranged response options horizontally - We included instructions for rating overall credibility - We experimented with short versions of ICEMAN but decided to stick to long version only - We discussed how we should frame the additional consideration – negative, positive, or offer both. We decided to offer only positive considerations. It reduced complexity and helps to make clear that credibility won't decrease if nothing applies. - Most discussion points were fine points and wording issues that are not shown here
1. A priori hypothesis (vs post-hoc explanation; confirmatory vs exploratory; hypothesis testing vs hypothesis generating); credibility is higher if the investigators stated a hypothesis prior to performing the study, lower if an explanation arose after data analysis. (n=41) ;	7, 7, 7, 7, 7, 7, 7, 7, 6; (6.8)	Definitely include	**Core item 1**: "Was the effect modification correctly hypothesized a priori?"		Someone pointed out that "correct", referring to correct direction, was not clear. Someone pointed out that we may penalize the honest if we make explicit statements of exploratory the worst category but unclear statements the next higher category. We decided to leave it as is because we see no other way without collapsing response options (in which case we would lose reliability). There were concerns about including the item in the meta-analysis version as people may construct a pseudo a priori hypothesis around known data (given that MAs are retrospective). We decided to keep the item to give credit to investigators who have a detailed protocol. We decided to mention thus potential limitation in the manual and await the results of the user testing.	No major revisions
2. Prior probability (strength of hypothesis); credibility increases with the prior probability of the effect modification being true (n=15)	7, 6, 4, 4, 4, 3, 2, 2, 1, 1; (3.2)	Combine items 2, 4, 5, and 6, e.g. framed as "extent of prior knowledge"	**Core item 2**: "Was the effect modification supported by prior evidence and theory?"		Someone pointed out that our explanation for the second response options (seen in some RCTs but not others) would be too strict and we should not expect the same effect modification in each and every RCT due to chance. We decided to delete this bit and elaborate on the issue in the manual. There were concerns about including the item in the meta-analysis version as people may pick prior evidence based on already know data (given that MAs are retrospective). We decided to delete the item for meta-analysis.	We decided to drop this core item in the MA version and make additional consideration worded "external evidence". This item seems less relevant for meta-analyses that are supposed to include all relevat RCTs and also are retrospective which makes definition of a priori difficult.
3. Bayesian analysis; Credibility is higher if priors were incorporated using explicit Bayesian methods (as opposed to informal methods) (n=5)	7, 6, 5, 4, 4, 4, 3, 2, 1, 1; (3.7)	Potential additional consideration	Excluded because too specific, idea covered by core item 1' and 2'			
4. Direction (specific vs. unspecific hypothesis); 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. (n=11)	7, 7, 7, 6, 5, 5, 4, 3, 1, 2; (5.0)	Combine items 2, 4, 5, and 6, e.g. framed as "extent of prior knowledge"	Covered by core item 1'			
5. Causal rationale; (biologic rationale, clinical rationale, mechanism); credibility is higher if there is a compelling causal model explaining the effect modification, and lower if not (n=42)	7, 7, 6, 5, 5, 5, 4, 4, 1, 7; (4.9)	Combine items 2, 4, 5, and 6, e.g. framed as "extent of prior knowledge"	Covered by core item 2'			
6. Indirect evidence (as opposed to prior theory only); credibility is higher if indirect evidence supports the effect modifier, e. g. evidence from animal studies, observational studies, related populations, or related interventions (n=17)	6, 6, 6, 6, 5, 5, 3, 3, 1, 7; (4.6)	Combine items 2, 4, 5, and 6, e.g. framed as "extent of prior knowledge"	Covered by core item 2'			
7. Expert input; credibility is higher if content expert were involved in the selection of candidate effect modifiers (n=4)	7, 5, 8, 5, 4, 2, 2, 2, 2, 1, 1; (3.1)	Definitely exclude from both versions of instruments	Excluded because too indirect, idea covered by core item 1' and 2'			
8. Consistent across studies (reproducibility, replicability, confirmation); credibility is higher if the effect modification is consistent across independent studies (n=28)	7, 7, 7, 7, 7, 6, 6, 6, 5, 1; (5.9)	Definitely include	**Core item 3': "If within-RCT comparisons are available, is the effect modification consistent across RCTs?"**	**Core item 4': "Is the effect modification consistent across previous RCTs?"**	For RCTs, we decided to combine with core item 2 ("Was the effect modification supported by prior evidence and theory ?")	We revised the response options and explanations now distinguishing between and within trial comparisons also for IPD meta-analysis
9. Consistent across outcomes; credibility is higher if the effect modification is consistent across related outcomes (n=6)	7, 6, 5, 5, 5, 3, 3, 3, 2, 1; (4.0)	Potential additional consideration	Included as additional consideration		No major revisions	No major revisions
10. Pre-specification of analytic details; credibility is higher if analytic details have been pre-specified prior to the analysis (n=31)	7, 7, 7, 7, 6, 6, 5, 5, 4, 1; (5.5)	Combine with item 11	Covered by core item 5'			
11. Multiplicity addressed; (inflated p-value, data-dredging); credibility is higher if investigators accounted formally or informally for multiplicity (n=39)	7, 7, 7, 7, 6, 6, 6, 4, 2, 7; (5.8)	Definitely include	**Core item 5': "Are multiplicity issues unlikely?"**		We extensively discussed the multiplicity item and decided to frame it positively (are multiplicity issues adequately dealt with?); We came to consensus to specify number of 1 of indicting low risk of multiplicity issues; we will acknowledge in the manual that the number is arbitrary but based on expert consensus; we will give credit if investigators adjust for multiplicity (which is rare); We decided to suggest to count effect modifiers (the variable) but explain in the manual that there may be additional layers of multiplicity such as cut points, time points, outcomes; We included a footnote clarifying that the multiplicity item is about the reported results whether or not adjusted for multiplicity.	we simplified the explanation, moving additional sources of multiplicity to manual only
12. Complete reporting (vs incomplete or selective reporting); credibility is higher if all performed analyses and results are reported (n=23)	7, 7, 7, 6, 5, 5, 5, 5, 4, 3; (5.4)	Combine with item 11	Covered by core item 5'			
13. Small number of candidate effect modifiers; credibility is higher if only a small number of effect modifiers have been tested (n=31)	7, 7, 7, 7, 5, 4, 4, 3, 2, 1; (4.7),	Combine with item 11	Covered by core item 5'			
14. Primary outcome; credibility is higher for effect modifiers of primary rather than secondary outcomes (n=4)	7, 7, 7, 6, 6, 6, 6, 3, 2, 1; (5.1)	Combine with item 17	Excluded because too indirect and unreliable			
15. Interaction test (test of homogeneity, test of heterogeneity); credibility is higher if an interaction test suggests a small likelihood for a chance finding (n=45)	7, 7, 7, 7, 7, 7, 6, 6, 6, 7; (6.8)	Definitively include	**Core item 6: "Is chance a very unlikely explanation of the apparent effect modification?"**		We discussed whether the interaction test item should be a stop item (e.g. stop if the interaction p-value is ≤.8) but decided to keep is as a regular item. The group expressed concerns about both using p-values in general and suggesting cut points for p-values; other measures such as interaction confidence intervals or Bayes factors are more informative. Because those are almost never reported, we decided to leave p-values in the instrument and refer to better alternatives in the manual. When we discussed interaction p-values, members of the group referred to a recent paper (Benjamin 2018) that suggests a shift from 0.05 to 0.005 as a significant threshold in medical science. In the draft, we suggest <0.01 for the highest category. We discussed intensively whether we should shift the thresholds. We suggest to keep <0.01 and but adapted the wording (highest category not unlikely instead of very unlikely). We were concerned that we would lose responsiveness of the item and thus reliability of the overall rating. Achieving a definitely credible in the overall rating will still be very difficult.	Ongoing discussion regarding specification of p-values. Some members of the group would have preferred more strict p-values, some members were against more strict p-value due to concerns about responsiveness. As a compromise, we weakened the wording making "chance is unlikely" the best response option.
16. multivariable model (independent, adjusted, joint effect vs marginal effect); credibility is higher if the effect modification is independent from other effect modifications. (n=8)	6, 5, 5, 5, 4, 5, 4, 4, 3, 3, 3; (4.3)	No decision, core group suggested include as potential additional consideration, no objection	Included as additional consideration		No major revisions	No major revisions
17. Power (type 2 error); credibility increases with the power to detect the effect modifier (n=20)	6, 6, 6, 6, 6, 4, 2, 1, 1, 1; (4.0)	No decision, core group suggested include as potential additional consideration, no objection	Included as additional consideration		We added an statement anchored to a number suggesting that high power (e.g. 80%) would be exceptional for an analysis of effect modification	We failed to find a convincing example for this additional consideration
18. Risk of bias overall effect; credibility is lower if the overall treatment effect is at risk of bias. (n=5)	7, 6, 6, 5, 5, 4, 3, 3, 1, 1; (4.1),	Definitely exclude from both versions of instruments	Included as additional consideration		We included suggestions for risk of bias instruments	No major revisions
19. Prognostic balance within subgroups; credibility is higher if, within each subgroup, prognostic factors are balanced (n=5)	5, 4, 4, 4, 5, 1, 1, 1, 1, 7; (2.9)	Potential additional consideration	Included as additional consideration		We dropped this consideration because to complex	
20. Stratified randomization; credibility is higher if the effect modifier was used as a stratification variable at randomization. (n=12)	7, 7, 5, 4, 3, 3, 3, 2, 1, 1; (3.6),	Definitely exclude from both versions of instruments	Excluded because not directly relevant for credibility of claimed effect modification			
21. Sample size per subgroup; credibility is higher if the sample size is large and balanced across subgroups under consideration (n=11)	7, 7, 7, 6, 4, 3, 3, 1, 1, 7; (4.3)	Definitely include (meta-analysis only, then number of studies per subgroup)	**Core item 7': "If between-RCT comparison, is the number of studies large?"**	Not applicable to RCTs	We came to consensus to specify a number, acknowledging in the manual that is arbitrary. We decided that 10 studies in the smallest subgroup is sufficiently strict for the highest category. We decided to discuss meta-regression in the manual (to do) because providing a single number (e.g. 20 studies) can be misleading if the distribution of studies is not balanced across levels of the effect modifier.	Because of concerns of responsiveness, we decided to reduce the minimum number of studies in the smallest subgroup for the highest response option.
22. Large effect modification (magnitude); credibility is higher if the effect modification is large. (n=5)	7, 6, 6, 5, 5, 3, 3, 2, 1, 1; (3.9)	Definitely exclude from both versions of instruments	Excluded because ambiguous, some experts felt large effect modification is reassuring, others were sceptical of large effects. Better covered by core item 6'.			We discussed again whether we should include the point estimate of effect modification with confidence intervals rather than interaction p-values only but decided to leave out estimation. The currently appear
23. 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 (n=22)	7, 7, 7, 7, 6, 5, 4, 3, 3, 1; (5.0)	Restrict instrument to effect modifiers measured at randomisation; include as preliminary consideration	Included as preliminary consideration as a stop-item		No major revisions	We refined the preliminary consideration because some post-randomization variables such as risk of bias may be appropriate as effect modifiers as long as considered on the study level
24. Confounding (marker, surrogate, proxy, common cause); credibility is higher if the effect modifier is a cause of the outcome as opposed to being a proxy for an underlying cause (e.g. sex might be a proxy for age). (n=19)	7, 6, 5, 5, 4, 3, 3, 2, 1, 7; (4.0)	No decision, core group suggested exclusion, no objection	Excluded because to complex and rarely possible. Causality is not necessary for a credible effect modification			
25. Within versus between study comparison (individual's vs study level, individual patient data vs aggregate data); credibility is higher if the candidate effect modifier was identified in individual participant data (within-study comparison) rather than study-level data (between study comparison) (n=22)	7, 7, 7, 7, 7, 5, 5, 3, 1; (5.4)	Definitely include (meta-analysis only)	**Core item 8': Is the effect modification based on within rather than between study comparison?**	Not applicable to RCTs	No major revisions	We refined the item making clear that IPD meta-analysis does not guarantee a within-study comparison
26. Measurement error; credibility is higher if the effect modifier is measured without error (e.g. no misclassification) (n=13)	6, 5, 4, 3, 3, 3, 2, 2, 1; (3.3)	Definitely exclude from both versions of instruments	Excluded because measurement error likely dilutes an effect modification			
27. Categorization (vs continuous); credibility is higher if continuous outcomes are not categorized (n=5)	7, 7, 6, 4, 4, 4, 3, 3, 3, 2; (4.3)	Combine items 27, 28, 29, 30 in item addressing issues of continuous effect modifiers	**Core item 9': "If the effect modifier was continuous, was it analyzed appropriately?"**		No major revisions	We focussed the item on issues of cut point selection and moved more complex considerations about continuous effect modification to the manual
28. Unusual threshold; credibility is lower if a threshold is unnatural and not justified (n=1)	7, 6, 5, 5, 4, 3, 2, 2, 1, 7; (3.9)	Combine items 27, 28, 29, 30 in item addressing issues of continuous effect modifiers	Covered by core item 9'			
29. Model specification considered for continuous effect modifiers (non-linearity considered); credibility is higher if the researchers considered the appropriateness of the functional form of the model for continuous effect modification, such as non-linear relationships (n=11)	7, 5, 5, 5, 4, 4, 4, 2, 1, 7; (4.1)	Combine items 27, 28, 29, 30 in item addressing issues of continuous effect modifiers	Covered by core item 9'			
30. Dose-response relationship; credibility is higher if there is a dose-response relationship across ordered levels of an effect modifier (including continuous) (n=3)	7, 6, 6, 6, 4, 4, 3, 3, 3, 7; (4.7)	Combine items 27, 28, 29, 30 in item addressing issues of continuous effect modifiers	Included as additional consideration		No major revisions	No major revisions
31. Quantitative vs qualitative; credibility is higher if the effect modification is quantitative (direction of effect consistent across levels of the effect modifier) rather than qualitative (direction varies by levels of the effect modifier). (n=12)	6, 5, 5, 4, 3, 3, 1, 1, 1, 7; (3.2)	Definitely exclude from both versions of instruments	Excluded, no general rule seemed possible			
32. Random effects model; credibility is higher if the analysis allows for true variation between studies within a subgroup. (n=5)	7, 7, 6, 4, 4, 4, 4, 3, 1; (4.4)	Definitely include (meta-analysis only)	**Core item 10': "If between-RCT comparison, was the analysis based on a random-effects rather than fixed effects model?"**	Not applicable to RCTs	No major revisions	We clarified terminology in the manual (i.e. common effect, fixed effects, random effects, mixed effects)
33. Regression on control group risk; credibility is lower if the effect modifier is the control group risk of an outcome. (n=0)	6, 6, 5, 4, 4, 4, 3, 3, 1, 1; (3.7)	Definitely exclude from both versions of instruments	Covered by preliminary consideration that excludes effect modifiers measured after randomisation			
34. Sensitivity analysis; credibility is higher if a sensitivity analysis suggests robustness to relevant assumptions such as thresholds or type of model (n=6)	7, 7, 6, 6, 5, 4, 3, 3, 1; (4.8)	Potential additional consideration	Included as additional consideration			
35. Significant overall effect (positive vs negative trial); credibility is higher if the overall effect is statistically significant (n=16)	6, 5, 5, 4, 4, 3, 2, 1, 1, 7; (3.4)	Potential additional consideration	Included as additional consideration		No major revisions	Dropped as additional consideration because the item was too controversial
36. Shrinkage applied; credibility is higher if investigators used shrinkage methods (weighted average of overall effect and subgroup specific effect) to adjust for optimism (n=3)	7, 7, 6, 5, 5, 4, 4, 4, 2, 1; (4.5)	Potential additional consideration	Covered by core item 5'		No major revisions	Mention in manual only

Appendix 2: Summary of the input from the user testing

Interview Question	Number of users who made one or more positive comments	Number of users who made one or more critical comments or suggestions	Critical comments or suggestions that led to changes	Critical comments or suggestions for which we judged changes are not necessary
What was your overall experience when you applied ICEMAN?	14	1	"When you are using it for the first time, it will take a lot of time"; simplify the manual, revise structure, improved navigation; include key instructions in main sheet	None
How happy or unhappy would you feel about using the instrument again?	13	1	None	"As a journal editor, I hope I don't have to personally do it for every paper. I hope it would be my statistical reviewer who would use it or the regular reviewers with more statistical background"
Are there any additional items that you would include?	1	3	Mention in limitations that appropriateness of model is beyond scope of instrument	"If there is a negative trial, people are doing many subgroup analyses and then they find something and focus on it and it is likely not very credible" (we had decided in the consensus study to not include overall significance as a credibility item); "qualitative effect modification is a sign for less credible. I don't know if it could be an item to be honest, but it was what I was thinking may be a gap" (we had decided in the consensus study to not include qualitative effect modification as a credibility item)
What do you think about … … the number of questions?	13	0	None	None
… the structure of the instrument?	6	0	None	None
… the language of the instrument?	4	5	"Effect modification put me off"; replace difficult words such as putative, spurious, multiplicity, consistent, model; include clarification of terminology at top of sheet	Translate in other languages (perhaps at later time point)

Appendix 2 (continued)

… the format and layout	8	5	Make the instrument sheet an editable word file instead of fillable form which causes software compatibility problems; make comment section more visible; change font size; provided more space for preliminary considerations	Instrument doesn't fit one page which complicated the overall assessment. We experimented with short versions but eventually decided to stick with one full version; font still small if printed; We will create a excel template that can be used for data collection, e.g. for meta-research
… the name "ICEMAN"	6	1	None	"Would be nice if the meaning was related to effect modification"
What was your experience in applying … … the preliminary considerations	7	9	Improve wording; improve format; add protocol reference; include examples; add time-point; make clear that one form applies only to a single effect modification	Clarify that "measured before randomization" refers to the original RCTs
… the item: Was the direction of effect modification correctly hypothesized a priori?	8	5	Mention statistical analysis plan as an alternative to protocol; Include preliminary consideration addressing availability of study protocol; improve instruction in manual	None
… the item: Was the direction of effect modification correctly hypothesized a priori? (RCT only)	4	2	Include "no evidence" in second response option; improved example	None
… the item: Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification?	8	6	Revise negative wording (was "support against chance"); add explanation; revise cut-points for p-values that used wording "like around 0.05 and around 0.01 … e.g. .03 what is that? Around is such a vague word … probably too vague if there are multiple raters"; add to preliminary considerations that clinical relevance is not part of the credibility assessment; mention in limitations that appropriateness of model underlying test is too sophisticated;	Difficult to understand for some users "there is not a lot of understanding about chance"

Appendix 2 (continued)

the item: Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis??	5	9	Replace difficult words adjusted and multiplicity; Simplify explanation and mention only number of effect modifiers and potential adjustment, but not other potential sources of multiplicity	Comment to latest version: "I think it is far too academic for most clinicians"
… the item: If the effect modifier is a continuous variable, were arbitrary cut points avoided?	9	2	None	Explain what continuous means "I instantly understood what a continuous variable is but not everyone would"
… the item: Is the analysis of effect modification based on comparison within rather than between trials? (meta-analysis only)	4	2	Revise wording; improve explanation and examples	Difficult to understand for some users. "So you understand now why I was saying you need some statistical background … and this is the first question!"
… the item: For within-trial comparisons, is the effect modification similar from trial to trial? (meta-analysis only)	2	3	Replace "consistency across studies", which was unclear with "similar from study to study" and improve explanation in manual	None
… the item: For between-trial comparisons, is the number of trials large? (meta-analysis only)	4	0	None	None
… the item: Did the authors use a random effects model? (meta-analysis only)	5	2	Simplify the question and the explanations, remove "mixed effects" which flummoxed users	None
… the optional item: Are there any additional considerations that may increase or decrease credibility?	2	8	Include word "optional"; simplify the item by removing a checklist of additional considerations; include link to manual; allow both increase and decrease credibility; make two response options corresponding to probably reduced or increased	None
… the last item: How would you rate the overall credibility of the proposed effect modification?	7	13	"The hardest thing would be filling in the overall credibility … where to place that cross?" Revise labels (users didn't know where to put uncertain: we modified the concept from probability of being true to strength of evidence, then infer probability of being true in second step through explanation); add explicit instructions; include interpretation; include consequences; improve manual navigation;	Some users prefer categories; some prefer the continuous option

ICEMAN – instrument for assessing the credibility of effect modification analyses
RCT – randomized controlled trial

Appendix 3: Instrument for assessing the credibility of effect modification analyses (ICEMAN) in a randomized controlled trial

Quick Instructions

- Synonyms for effect modification include subgroup effect, interaction, and moderation
- The instrument applies to a *single* proposed effect modification at a time; complete one form per each outcome, time-point, effect measure, and effect modifier
- Response options on the left indicate definitely or probably reduced, response options on the right probably or definitely increased credibility
- Completely unclear goes under probably reduced credibility
- It is helpful to provide a supporting comment or quotation under each question
- Whether an effect modification is patient-important is *not* part of the credibility assessment
- The manual provides more detailed instructions and examples

Preliminary considerations

Study reference(s):

If available, protocol reference(s):

State a single outcome and, if applicable, time-point of interest (e.g. mortality at 1 year follow-up):

State a single effect measure of interest (e.g. relative or absolute risk difference):

State a single potential effect modifier of interest (e.g. age or comorbidity):

Was the potential effect modifier measured before randomization? [ ] yes, continue [ ] no, stop here and refer to manual for further instructions

Credibility assessment

1: Was the direction of the effect modification correctly hypothesized a priori?

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Clearly post-hoc or results inconsistent with hypothesized direction or biologically very implausible	*Vague hypothesis or hypothesized direction unclear*	*No prior protocol available but unequivocal statement of a priori hypothesis with correct direction of effect modification*	*Prior protocol available and includes correct specification of direction of effect modification, e.g. based on a biologic rationale*

Comment:

2: Was the effect modification supported by prior evidence?

[ ] Inconsistent with prior evidence	[ ] Little or no support or unclear	[ ] Some support	[ ] Strong support
Prior evidence suggested different direction of effect modification	*No prior evidence or consistent with weak or very indirect prior evidence (e.g. animal study or at high risk of bias) or unclear*	*Consistent with more limited or indirect prior evidence (e.g. large observational study, non-significant effect modification in prior RCT, or different population)*	*Consistent with strong prior evidence directly applicable to the clinical scenario (e.g. significant effect modification in related RCT)*

Comment:

3: Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification? (consider irrespective of number of effect modifiers)

[ ] Chance a very likely explanation	[ ] Chance a likely explanation or unclear	[ ] Chance may not explain	[ ] Chance an unlikely explanation
Interaction p-value > 0.05	*Interaction p-value ≤ 0.05 and > 0.01, or no test of interaction reported and not computable*	*Interaction p-value ≤ 0.01 and > 0.005*	*Interaction p-value ≤ 0.005*

Comment:

4: Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Explicitly exploratory analysis or large number of effect modifiers tested (e.g. greater than 10) and tested and number not considered in analysis multiplicity not considered in analysis	*No mention of number or 4-10 effect modifiers tested and number not considered in analysis*	*No protocol available but unequivocal statement of 3 or fewer effect modifiers tested*	*Protocol available and 3 or fewer effect modifiers tested or number considered in analysis*

Comment:

5: If the effect modifier is a continuous variable, were arbitrary cut points avoided? [ ] not applicable: not continuous

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Analysis based on exploratory cut point (e.g. picking cut point associated with highest interaction p-value)	*Analysis based on cut point(s) of unclear origin*	*Analysis based on pre-specified cut points, e.g. suggested by prior RCT*	*Analysis based on the full continuum, e.g. assuming a linear or logarithmic relationship*

Comment:

6 Optional: Are there any additional considerations that may increase or decrease credibility? (manual section 3.6)

	[ ] yes, probably decrease credibility	[ ] yes, probably increase credibility

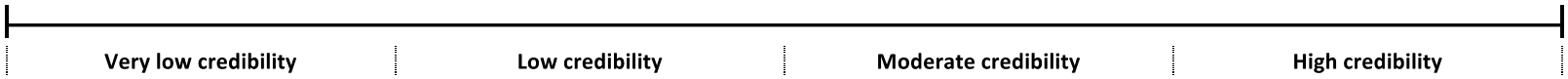
Comment:

7: How would you rate the overall credibility of the proposed effect modification?

The overall rating should be driven by the items that decrease credibility. The following provides a sensible strategy:

- All responses definitely or probably reduced credibility or unclear → very low credibility
- Two or more responses definitely reduced credibility → maximum usually low credibility even if all other responses satisfy credibility criteria
- One response definitely reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- Two responses probably reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- No response options definitely or probably reduced credibility → high credibility very likely

Place a mark on the continuous line (e.g. hit "x" in electronic version)



Very low credibility	**Low credibility**	**Moderate credibility**	**High credibility**
Very likely no effect modification Use overall effect for each subgroup	Likely no effect modification Use overall effect for each subgroup but note remaining uncertainty	Likely effect modification Use separate effects for each subgroup but note remaining uncertainty	Very likely effect modification Use separate effects for each subgroup

Comment:

Appendix 4: Instrument for assessing the credibility of effect modification analyses (ICEMAN) in a meta-analysis of randomized controlled trials

Quick instructions

- Synonyms for effect modification include subgroup effect, interaction, and moderation
- The instrument applies to a *single* proposed effect modification at a time; complete one form per each outcome, time-point, effect measure, and effect modifier
- Response options on the left indicate definitely or probably reduced, response options on the right probably or definitely increased credibility
- Completely unclear goes under probably reduced credibility
- It is helpful to provide a supporting comment or quotation under each question
- Whether an effect modification is patient-important is *not* part of the credibility assessment
- The manual provides more detailed instructions and examples

Preliminary considerations

Study reference(s):

If available, protocol reference(s):

State a single outcome and, if applicable, time-point of interest (e.g. mortality at 1 year follow-up):

State a single effect measure of interest (e.g. relative or absolute risk difference):

State a single potential effect modifier of interest (e.g. age or comorbidity):

Was the potential effect modifier measured before randomization? [ ] yes, continue [ ] no, stop here and refer to manual for further instructions

Credibility Assessment

1: Is the analysis of effect modification based on comparisons within rather than between trials?

[ ] Completely between	[ ] Mostly between or unclear	[ ] Mostly within	[ ] Completely within
Subgroup analysis or meta-regression comparing overall effects of each individual trial. This is typical for aggregate data meta-analysis.	*Subgroup analysis or meta-regression with most information coming from overall effects, but some trials providing within-trial subgroup information*	*Most trials providing within-trial subgroup information; or individual participant data analysis that combines within and between trial information*	*Individual participant data analysis that separates within from between trial information, e.g. meta-analysis of interactions*

Comment:

2: For within-trial comparisons, is the effect modification similar from trial to trial? [ ] Not applicable: no or one within-RCT comparison

[ ] Definitely not similar	[ ] Probably not similar or unclear	[ ] Mostly similar	[ ] Definitely similar
Effect modification reported for two or more trials and clearly different directions	*Effect modification not reported for individual trials or too imprecise to tell*	*Effect modification reported for two or more trials, mostly similar in direction, but considerable differences in magnitude*	*Effect modification reported for two or more trials, similar in direction, only some differences in magnitude*

Comment:

3: For between-trial comparisons, is the number of trials large? [ ] Not applicable: no between RCT comparison

[ ] Very small	[ ] Rather small or unclear	[ ] Rather large	[ ] Large
1 or 2 in smallest subgroup; 5 or less in continuous meta-regression	*3-4 in smallest subgroup; 6-10 in continuous meta-regression*	*5-9 in smallest subgroup; 11 to 15 in continuous meta-regression*	*10 or more in smallest subgroup; more than 15 in continuous meta-regression*

Comment:

4: Was the direction of effect modification correctly hypothesized a priori?

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Clearly post-hoc or results inconsistent with hypothesized direction or biologically very implausible	*Vague hypothesis or hypothesized direction unclear*	*No prior protocol available but unequivocal statement of a priori hypothesis with correct direction of effect modification*	*Prior protocol available and includes correct specification of direction of effect modification, e.g. based on biologic rationale*

Comment:

5: Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification? (consider irrespective of number of effect modifiers)

[ ] Chance a very likely explanation	[ ] Chance a likely explanation or unclear	[ ] Chance may not explain	[ ] Chance an unlikely explanation
Interaction or meta-regression p-value > 0.05	*Interaction or meta-regression p-value < 0.05 and > 0.01, or no test of interaction reported and not computable*	*Interaction or meta-regression p-value < 0.01 and > 0.005*	*Interaction or meta-regression p-value < 0.005*

Comment:

6: Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Explicitly exploratory analysis or large number of effect modifiers tested (e.g. greater than 10) and multiplicity not considered in analysis	*No mention of number or 4-10 effect modifiers tested and number not considered in analysis*	*No protocol available but unequivocal statement of 3 or fewer effect modifiers tested*	*Protocol available and 3 or fewer effect modifiers tested or number considered in analysis*

Comment:

7: Did the authors use a random effects model?

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Fixed (or common) effect(s) explicitly stated	*Probably fixed (or common) effect(s)*	*Probably random (or mixed) effects*	*Random (or mixed) effects explicitly stated*

Comment:

8: If the effect modifier is a continuous variable, were arbitrary cut points avoided? [ ] not applicable: not continuous

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Analysis based on exploratory cut point(s), e.g. picking cut point associated with highest interaction p-value	*Analysis based on cut point(s) of unclear origin*	*Analysis based on pre-specified cut point(s), e.g. suggested by prior RCT*	*Analysis based on the full continuum, e.g. assuming a linear or logarithmic relationship*

Comment:

9 Optional: Are there any additional considerations that may increase or decrease credibility? (manual section 4.9)

[ ] yes, probably decrease credibility [ ] yes, probably increase credibility

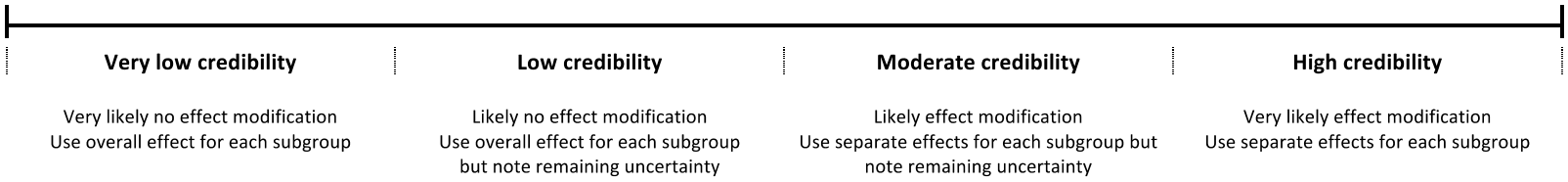
Comment:

10: How would you rate the overall credibility of the proposed effect modification?

The overall rating should be driven by the items that decrease credibility. The following provides a sensible strategy:

- All responses definitely or probably decrease credibility or unclear → very low credibility
- Two or more responses definitely decrease credibility → maximum usually low credibility even if all other responses satisfy credibility criteria
- One response definitely decreases credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- Two responses probably decrease credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- No response options definitely or probably decrease credibility → high credibility very likely

Place a mark on the continuous line (e.g. hit "x" in electronic version)



Comment:

Appendix 5: Instrument for the credibility of effect modification analyses (ICEMAN) in randomized controlled trials and meta-analyses: explanation and elaboration

Stefan Schandelmaier, Matthias Briel, Ravi Varadhan, Christopher H Schmid, Niveditha Devasenapathy, Rodney A Hayward, Joel Gagnier, Michael Borenstein, Geert JMG van der Heijden, Issa J Dahabreh, Xin Sun, Willi Sauerbrei, Michael Walsh, John PA Ioannidis, Lehana Thabane, Gordon H Guyatt

Table of contents

1 Introduction

The purpose of ICEMAN is to assess the credibility of an effect modification proposed in a randomized controlled trial (RCT) or meta-analysis of randomized controlled trials.

Effect modification means that the effect of an intervention on an outcome varies depending on another variable, called an effect modifier, such as age or comorbidity.[1]

Authors use various terminology for effect modification including subgroup effect, treatment-covariate interaction, moderation, or heterogeneity of treatment effects. Alternative names for effect modifier include moderator or predictive factor.

ICEMAN applies to **claims that an effect modification is present** (not to claims that an effect modification is absent – such an instrument would require different criteria).

The instrument can be used both by **authors of RCTs or meta-analyses** who are considering to make a claim of effect modification and by **individuals who are critically appraising an effect modification** claimed in a publication (e.g. systematic review authors, health technology assessment practitioners, guideline developers, journal editors, reviewers of journal articles, health care system decision makers, and health care providers).

The assessment starts with a set of preliminary considerations to clarify the sources of information and help define the effect modification under consideration (section 2 of this manual).

ICEMAN includes five core credibility considerations for RCTs (section 3) and eight core credibility considerations for meta-analyses (section 4). The considerations are framed as signaling questions, each with four response options. The response options differ in wording but share the same order and logic: Response options on the left indicate reduced credibility, response options on the right increased credibility. Unclear is combined with probably decreased credibility.

One optional question allows additional credibility considerations. Because additional considerations should have less influence on overall credibility than the core questions, the response options include only probably increased or decreased credibility. Leaving the question blank does not affect credibility.

The final part is an overall credibility rating based on the credibility considerations expressed on a continuous scale divided into four credibility areas.

This manual provides, for each element of ICEMAN, more detailed explanations, a rationale with key references for support, examples from the literature, and an example of a completed instrument for both RCTs and meta-analyses.

Section 5 explains how ICEMAN works in context and can be combined with risk of bias tools and GRADE.

Section 6 includes more in-depth conceptual explanations for interested readers.

2 Preliminary considerations

Study reference(s): Use this section to provide a link to the study or publication(s) under consideration. It may also be helpful to specify the comparison of interest, especially if a study includes more than two arms.

If available, protocol reference(s): For optimal assessment, some credibility considerations require that the authors have produced an accessible study protocol or statistical analysis plan, ideally time-stamped. If available, provide a link to a study protocol (e.g. a published protocol or an entry in a study registry such as ClinicalTrials.gov). Many protocols, however, provide insufficient information regarding analyses of effect modification.

State a single outcome and, if applicable, time-point of interest: Use this section to specify a single outcome of interest. In most studies, there is only one population, intervention, and comparator, but usually multiple outcomes. Because ICEMAN refers to a single outcome at a time, users must specify the outcome of interest and, if applicable, the time-point of outcome assessment (e.g. mortality at 1 year follow-up).

State a single effect measure of interest: Use this section to specify a single effect measure of interest (e.g. relative risk, risk difference, odds ratio, or hazard ratio for binary outcomes, or difference or ratio of means for continuous outcomes).

The type of effect measure is a key consideration because the magnitude of effect modification typically differs between effect measures, and in particular between measures of relative versus absolute effect.[2-4] Therefore, the credibility rating is likely to differ depending on the chosen effect measure.

State a single potential effect modifier of interest (e.g. age, comorbidity): Use this section to specify the potential effect modifier of interest (i.e. one effect modifier on each ICEMAN form). Effect modifiers may be patient characteristics (e.g. disease severity, age, or body mass index), intervention alternatives (e.g. dose, co-interventions, mode of administration), or methodological study characteristics (e.g. risk of bias, outcome definition, type of funding). The instrument does not apply when the effect modifier is another outcome (see following section).

For continuous effect modifiers, it may also be helpful to specify any thresholds used.

Note that an effect modifier (e.g. sex) is different from a particular subgroup (e.g. women).

Was the effect modifier measured before randomization? If the effect modifier is another, typically intermediate, outcome, the assessment of effect modification is complicated and potentially misleading.[5-19] Those analyses require different methods [11,15,20] and result in less secure conclusions.

There are exceptions in which the instrument does apply to effect modifiers measured after randomization: 1) The effect modifier is a non-modifiable characteristic, e.g. sex or age; 2) For meta-analyses: the effect modifier is a study characteristic such as risk of bias, length of follow-up, or mean received dose.

> *Example: An RCT testing strict or conventional management of hyperglycemia with insulin therapy in ICU patients claimed an effect modification by length of hospital stay. Among patients who stayed in the ICU for less than three days, mortality was greater among those receiving intensive insulin therapy. In contrast, among patients who stayed in the ICU for three or more days, mortality was lower among those receiving intensive insulin.[21] Length of ICU stay, however, was shortened by the intervention. Therefore, the control patients needed a better baseline prognosis in order to qualify for the short-stay subgroup than patients in the intervention group. This prognostic imbalance between intervention and control group within the length of stay subgroups likely created the differences in mortality.*

3 ICEMAN for randomized controlled trials

3.1 Was the direction of effect modification correctly hypothesized a priori?

Item explanation: Credibility is higher if investigators correctly anticipated the direction of the effect modification (e.g. that an intervention is more effective in younger than in older patients), lower if they failed to anticipate a direction, and lowest if they anticipated the opposite direction. This item captures a number of credibility considerations:

Correct anticipation of an effect modification implies that the investigators had a specific hypothesis in mind – usually based on a biologic or other causal rationale, or sometimes based on prior evidence (see next item). For instance, investigators may have anticipated a stronger relative effect in younger than in older patients because a disease may be too advanced in older patients for the intervention to be effective. If the data conforms to this hypothesis, the credibility is increased, otherwise decreased.

If the a priori specification of the effect modification hypothesis does not include a direction (e.g. by specifying the in the protocol that that the effect may vary by age but failure to say whether the effect is greater in the old versus the young or the other way round) this is weaker and probably not much better than having no prior hypothesis at all. In the Bayesian framework, the idea of a specific a priori hypothesis corresponds to using an informative rather than non-informative prior.[22,23]

In addition, the item captures that an explanation (e.g. a biological rationale) stated a priori is much more credible than a post hoc explanation. If post hoc, investigators had likely considered many possible explanations, thereby creating a multiplicity problem.[6,9,24-28]

Moreover, the item captures that hypotheses are most credible if verified in a prior, ideally in a time-stamped protocol or analysis plan. Note that a statements of pre-specification may not be reliable,[29] nor do they imply a specific hypothesis.

Note that if an effect modifier was a stratification factor at randomization, it does not necessarily imply a specific hypothesis, but it may increase the likelihood that this was the case.

Note that the direction of an effect modification may depend on the type of effect measure if the effect modifier is also a prognostic factor (most are).
Response options and examples:

[ ] Definitely no: Clearly post-hoc or results inconsistent with hypothesized direction or biologically very implausible
Example 1: The ISIS-2 trial testing ASA conducted a provocative post hoc subgroup analysis comparing patients born under different astrological signs. ASA had a slight adverse effect in patients born under the sign of Gemini or Libra but a substantial benefit in patients born under other astrological signs. The example became famous because the finding was obviously post hoc and not compatible with any biological model.
Example 2: In a trial comparing the effect of vasopressin versus norepinephrine for septic shock on mortality, the authors had hypothesized that the benefit of vasopressin over norepinephrine would be larger in patients with more severe septic shock. It turned out, however, that the benefit of vasopressin seemed to be greater in the patients with less severe septic shock (RR 1.04 in more severe v 0.74 in less severe septic shock, interaction P=0.10). The investigators' failure to correctly identify the direction of the subgroup effect appreciably weakens any inference that vasopressin is superior to norepinephrine in the less severely ill patients.[30]

[ ] Probably no: Vague hypothesis or hypothesized direction unclear
Example: The investigators of the first large trial of aspirin for patients with transient ischemic attacks reported that aspirin had a beneficial effect in preventing stroke in men, but not in women with cerebrovascular disease.[31] For many years, this led many physicians to withhold aspirin from women with cerebrovascular disease. Although the investors may have planned a priori to explore subgroup effects by sex, they had not anticipated a specific direction based on biologic rationale or prior evidence. Therefore, the effect modification had a very low prior probability of the effect modification being true. Subsequent studies and meta-analyses failed to replicate the subgroup effect.[32]

[ ] Probably yes: No protocol available but unequivocal statement of a priori hypothesis with correct direction of effect modification.

Example: A trial in patients requiring dialysis compared jugular versus femoral catheterization and found no significant difference with respect to catheter colonization. An analysis of effect modification suggested that jugular catheters were superior in patients with high BMI but inferior in patients with a low BMI. The authors claimed that they had correctly anticipated the direction of the effect modification but there was not protocol available.[33]

[ ] Definitely yes: Prior protocol available and includes hypothesis with correct specification of direction of effect modification, e.g. based on biologic rationale
Example: A trial comparing two types of nails (reamed versus undreamed Intramedullary nails) for tibial shaft fractures suggested that reamed nails were superior for closed but potentially inferior for open fractures.[34] *The investigators correctly anticipated the direction of effect modification in their published protocol based on a biologic rationale: damage of endosteal blood supply through reamed nails may be more detrimental in open than in closed fractures.*[35]

3.2 Was the effect modification supported by prior evidence?

Item explanation: Credibility is higher if the effect modification is supported by prior direct or indirect evidence, lower if observed for the first time (often labelled as *exploratory*), lowest if inconsistent with prior evidence.

Replication, ideally in another RCT, makes chance a less likely explanation for an apparent effect modification. Attempts for replication and successful replication, however, seem to be rare[36] and prior evidence will be mostly unclear.

Similar to the previous item, direction plays an important role. If two trials show different directions of effect modification, this markedly reduces credibility. Because of the role of chance, however, we should not expect to see the same magnitude of effect modification in all trials. Many trials will be underpowered and some may show the opposite direction due to chance alone.

Response options and examples:

[ ] Inconsistent with prior evidence: Prior evidence suggested different direction of effect modification
[still looking for good example]

[ ] Little or no support or unclear: No prior evidence or consistent with weak or very indirect prior evidence (e.g. animal study at high risk of bias) or unclear
Example: A trial in patients requiring dialysis compared jugular versus femoral catheterization and found no significant difference with respect to catheter colonization.[33] *An analysis of effect modification suggested that jugular catheters were superior in patients with high BMI but inferior in patients with a low BMI. The authors claimed that they had correctly anticipated the direction of the effect modification and provided a reference to a previous cohort study. The prior evidence, however, was*

unclear because the cited study provided no direct support for an interaction and was published after the trial had already started.[37]

[ ] Some support: Consistent with more limited or indirect prior evidence (e.g. large observational study, non-significant effect modification in prior RCT, or different population)
Example: A trial testing Epoetin Alfa for critically ill patients suggested a reduction in mortality in patients who had a trauma but not in other patients.[38] *Although the interaction test was not significant (p-value 0.16), it was consistent with a similar effect modification seen in a previous RCT which – although not significant either – provides some empirical support.*[39]

[ ] Strong support: Consistent with strong prior evidence directly applicable to the clinical scenario (e.g. significant effect modification in related RCT).
Example: A trial comparing low-carb versus low-fat diet found suggested modification by amount of insulin secretion. Low-carb diet was superior in patients with high insulin secretion but inferior in patients with low insulin secretion (interaction p=0.01).[40] *A previous RCT cited in the paper found a similar, significant effect modification (interaction p=0.02).*[41]

3.3 Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification?

Item explanation: Credibility is higher if statistical test for interaction suggests that chance is an unlikely explanation for the apparent effect modification.[42-44] Credibility is lower if an interaction test suggests that an apparent effect modification is compatible with chance – or no test is available and impossible to compute. (Here we use the term interaction as a synonym for effect modification, acknowledging that some methodologists reserve the term for causal effect modification.[42])

For this item, consider the results of the interaction test (usually a p-value) as reported, irrespective of whether the p-value was adjusted for the number of analyses or not, or effect modifiers were analyzed jointly or one-by-one. We deal with considerations of multiple analyses separately in the following item.

Note that showing that an effect is significant in one subgroup and not in another is of little use: it provides no information whether chance might explain differences in effects across subgroups.[9,10,43,45,46]

A number of interaction tests are available. Most common in the context of RCTs is to include an interaction term in a regression model. Most reports of RCTs do not explicitly quantify the effect modification using a single number (e.g. by providing a ratio of risk ratios with associated

confidence interval). Instead, they typically provide a plot or a table showing subgroup-specific estimates, ideally accompanied by p-values from a test of interaction.

If no interaction p-value is reported, it can sometimes be calculated based on the reported data (point estimates of effect and confidence intervals in individual subgroups).[47,48] As rule of thumb is that the interaction p-value must be smaller than 0.05 if 95% confidence intervals of subgroup-specific estimates do not overlap.

We anchored the response options around typical thresholds for p-values 0.05, 0.01, and 0.005, with a p-value of 0.005 or smaller representing the most credible category. The response options recognize that p-value thresholds of 0.05 or even 0.01 may be too lenient for claiming statistical significance.[49] Of course, these are arbitrary settings and some methodologists would recommend even lower thresholds. Because of the low power of many analyses of effect modification, however, this would decrease the responsiveness of the item and the instrument's ability to distinguish more from less credible effect modification.

Note that other statistical measures than p-values such as interaction confidence intervals or Bayes factors may be more informative and intuitive than p-values but are rarely reported.

Response options and examples:

[ ] Chance a very likely explanation: Interaction p-value > 0.05
Example: A trial comparing prostatectomy versus observation for early prostate cancer found no difference after nearly 20 years of follow-up. Based on interaction p-values larger than 0.05, the investigators hypothesized a potential benefit in the subgroup of patients with a low PSA value (interaction p=0.06) and in the subgroup of patients with an intermediate risk tumor (interaction p=0.08).[50] The most likely explanation for those effect modifications is chance, especially considering the rating for the other items of the instrument.

[ ] Chance a likely explanation or unclear: Interaction p-value ≤ 0.05 and > 0.01, or no test of interaction reported and not computable
Example: The PLATO trial compared the two platelet inhibitors Ticagrelor and Clopidogrel regarding prevention of cardiovascular events. A subgroup analysis comparing patients from different continents suggested that Ticagrelor is superior in patients from all continents but North America where it seemed to be inferior (p=0.05). The p-value suggests that 1 in 20 trials may show such an effect modification or larger, even if not true.[51]

[ ] Chance may not explain: Interaction p-value ≤ 0.01 and > 0.005
Example: A trial comparing reamed versus unreamed Intramedullary nailing of tibial shaft fractures suggested that reamed nails a superior for open fractures but not for closed fractures. An interaction p-value of 0.01 provided modest support against chance.[34]

[ ] Chance an unlikely explanation: Interaction p-value ≤ 0.005
Example: In a trial testing the administration of tranexamic acid in bleeding trauma patients found that the effect on preventing death due to bleeding varied according to the time from injury to treatment. Early treatment ≤1 h from injury) significantly reduced the risk of deaths due to bleeding (relative risk 0.68), treatment given between 1 and 3 h also reduced the risk (RR 0.79), while treatment given after 3 h seemed to increase the risk of death due to bleeding (RR 1.44). The interaction p-value was smaller than 0.0001 suggesting that the apparent interaction is not a chance finding.[52]

3.4 Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

Item explanation: Performing multiple tests is a major concern in the context of effect modification analysis. Trialists usually measure a large number of baseline variables, many of which they could test for potential effect modification. Because multiple tests increase the risk of a chance finding,[53-55] credibility is higher if investigators have tested only a small number of effect modifiers. Conversely, credibility decreases with the number of tested candidate effect modifiers. We therefore advise counting the number of candidate effect modifiers stated, ideally verified in a protocol.

Multiplicity issues can arise in different ways.[56] Most obvious are situations in which investigators test multiple candidate effect modifiers and highlight significant results. Another important issue which we address in a separate item concerns selection of cut points of continuous effect modifiers. Other potential multiplicity issues include multiple time points, multiple scales,[57] multiple outcomes, or multiple methods for testing the interaction. Therefore, even if the number of effect modifiers is small, one should consider whether other issues might have introduced multiplicity.

An alternative to limiting the number of analyses is to statistically adjust the analysis for multiplicity. Credibility is higher if an effect modification persists after adjustment. Different techniques are available including correction of p-values considering the (familywise) type 1 error rate,[58] testing all candidate effect modifiers in a common model, using a composite variable such as a risk score, or shrinkage estimators.[45,59] All techniques inevitably reduce power.[6,60,61] Most, investigators, however, do not address potential multiplicity issues in design or analysis and leave the judgement to the reader – another reason why a small number of effect modifiers is most helpful.

Assessment of multiplicity crucially depends on reporting (reporting guidelines for effect modification are available [62-64]). Without knowing the number of effect modification analyses performed, we cannot assess the potential impact of multiplicity. Ideally, investigators would specify candidate effect modifiers along with definitions and analytic details in a protocol. If no protocol is available, one should look for explicit statements about the number of effect modifiers. A note of caution: an empirical study has shown that retrospective statements about

the number of pre-specified subgroup analyses are not always reliable.[29] Also note that a statement that a particular effect modifier was pre-specified does not rule out the problem of multiplicity because investigators may have pre-specified many other effect modifiers.

In summary, this item requires counting the number of effect modifiers (perhaps considering additional multiplicity issues), if possible verifying them in a protocol, and considering whether investigators considered the number of analyses in their statistical analysis.

Response options and examples:

[ ] Definitely no: Explicitly exploratory analysis or large number of analyses (e.g. greater than 10) and multiplicity not considered in analysis
Example: A trial assessing the risk of stroke after carotid endarterectomy for symptomatic stenosis suggested that the benefit of the surgery is reduced in patients taking only low-dose aspirin because of an increased operative risk.[65] The investigators had tested all their baseline factors (more than 20) for potential effect modification without adjustment for multiplicity. A subsequent trial randomizing patients undergoing endarterectomy to low and high dose aspirin suggested the opposite association, thus providing strong evidence against the claimed effect modification: benefit of endarterectomy was larger in the low dose group than in the high dose group.[66] Most likely, the multiple analysis in the first trial had identified a random result.

[ ] Probably no or unclear: No mention of number or 4-10 effect modifiers tested and number not considered in analysis
Example: A trial comparing prostatectomy versus observation for early prostate cancer found no difference after nearly 20 years of follow-up. Based on interaction p-values larger than 0.05, the investigators hypothesized a potential benefit in the subgroup of patients with a low PSA value (interaction p=0.06) and in the subgroup of patients with an intermediate risk tumor (interaction p=0.08).[50] In their interpretation, the investigators did not take into account that they had tested at least seven effect modifiers for this outcome.

[ ] Probably yes: No protocol available but unequivocal statement of 3 or fewer effect modifiers tested
Example: A trial tested whether training of health professionals reduces the number of cesarean deliveries. A subgroup analysis suggested that the intervention worked for women with low-risk pregnancies but not for women with high-risk pregnancies (interaction p=0.03). No a priori published protocol was available but investigators provided a protocol together with the main publication in which they explicitly pre-specified the two reported effect modifiers.[67]

[ ] Definitely yes: Protocol available and 3 or fewer effect modifiers tested or number considered in analysis
[still looking for good example]

3.5 If the effect modifier is a continuous variable, were arbitrary cut points avoided?

Item explanation: Categorizing continuous effect modifiers is common[68] but associated with a number of problems:[69,70] Cut points can introduce multiplicity, reduce power, mask linear or non-linear associations, and complicate comparisons across studies. Therefore, analyses that avoid cut points and make use of the full spectrum of values are the most credible.

Investigators often decide against using the complete data and rather use cut points to partition continuous effect modifiers in two or more categories. Categories with a strong, empirically grounded rationale, are the most credible. For instance, arbitrariness can be avoided by pre-specifying the cut points based on a previous RCT that demonstrates effect modification. Credibility is low if investigators selected the best-fitting data-driven cut point to maximize the effect modification. Such cut points are associated with a high rate of false positive claims.[69,71]

There are some challenges when modelling continuous effect modifiers that are not part of the instrument but may lower the credibility: model misspecification can occur if the continuous relationship is driven by a few influential observations.[72-74] Post-hoc modelling can lead to overfitting. Most credible are therefore continuous analyses for which investigators have pre-specified the type of dependency of the treatment effect on the continuous variable (sometimes referred to as *treatment effect function*) such as a linear or log relationship, or considered a small number of candidate functions.[75]

An alternative to use of cut points and potentially complex modelling is to consider overlapping subgroups (e.g. using a sliding window approach).[76] The credibility is usually much higher than using arbitrary cut points but the interpretation can be difficult.

The credibility of a continuous analysis usually increases if investigators present a plot with confidence bands around the regression function (often a line) and carefully checked the proposed model. Formal interaction tests for continuous effect modification are available and should be applied.[75]

Note that additional considerations related to continuous effect modifiers may apply, e.g. if there is a clear dose-response relationship or results were robust to sensitivity analyses (see following question).

Response options and examples:

> **[ ] Definitely no:** Analyzed based on exploratory cut point(s) (e.g. picking cut point associated with highest interaction p-value)
> *[still looking for good example]*
>
> **[ ] Probably no or unclear:** Analyzed based on cut point(s) of unclear origin

Example: A trial comparing prostatectomy versus observation for early prostate cancer found no difference after nearly 20 years of follow-up. Based on an interaction p-values of 0.06, the investigators hypothesized a potential effect modification by PSA value below or above 10. The investigator provided no rationale for the chosen threshold. A clear justification, an analysis based on the full spectrum, or a sensitivity analysis using different cut points or could have strengthened or discarded the hypothesized effect modification.[50]

[ ] Probably yes: Analysis based on pre-specified cut points, e.g. suggested by prior RCT
Example: In a trial testing the administration of tranexamic acid in bleeding trauma patients found that the effect on preventing death due to bleeding decreased with increasing time from injury (interaction p<0.0001). Early treatment ≤1 h from injury) significantly reduced the risk of deaths due to bleeding (relative risk 0.68), treatment given between 1 and 3 h also reduced the risk (RR 0.79), while treatment given after 3 h seemed to increase the risk of death due to bleeding (RR 1.44).[52] *The investigators had pre-specified the three categories in a published protocol.*[77]

[ ] Definitely yes: Analysis based on the full continuum, e.g. assuming a linear or logarithmic relationship.
Example: A trial comparing interferon-alpha versus medroxyprogesterone in patients with renal carcinoma found a benefit of interferon.[78] *A subsequent analysis suggested white cell count as an effect modifier: that the benefit of interferon – a toxic drug – seemed to disappear as the white cell count increased. The investigators treated white cell count as a continuous variable. By avoiding an arbitrary cut point, the investigators maximized the power of the analysis. A Plot of the continuous relationship provides confidence bands and shows a dose-response relationship.*[79]

3.6 Are there any additional considerations that may increase or decrease credibility?

Item explanation: Methodologists have suggested a number of additional considerations that could be relevant for assessing the credibility of effect modifiers (REF systematic survey, currently under review with JCE). They are not part of the core items because they either are less relevant, rarely apply, or are difficult to assess. Because they are usually less relevant than the previous core items, the only response options are probably decreased and probably increased.

Additional considerations are optional, that is, leaving this section blank does not affect credibility. Note that it may not be worth to consider potential additional considerations if core items already suggest low or very low credibility.

The following list provides potentially relevant additional considerations:

A sensitivity analysis suggested robustness to relevant assumptions: A sensitivity analysis can help to increase the confidence in a proposed effect modification.[15,80,81] For

instance, if an effect modification analysis is based on a categorized continuous variable, the credibility increases if the effect modification persists for different cut-points.
Example: In a trial testing the administration of tranexamic acid in bleeding trauma patients found that the effect on preventing death due to bleeding decreased with increasing time from injury (interaction p<0.0001). The authors used two cut points to define three subgroups (≤1 h from injury, between 1 and 3 h, and after 3 h). to assess the potential influence of choice of cut points, the authors performed a sensitivity analysis treating time as a continuous effect modifier which suggested that results were robust.[52]

"Dose-response effect" across levels of the effect modifier: Credibility may be higher if effects increase or decrease monotonically with increases in the levels of the modifier, e.g. an effect that increases incrementally across three or more age groups. On the contrary, it is especially important to beware of apparent effect modification that do not reflect a plausible pattern across three or more ordered groups, even if statistically significant. For instance, an effect might me abnormally elevated in one subgroup chosen from a continuum, but not in neighboring subgroups.
Example: In a trial testing the administration of tranexamic acid in bleeding trauma patients found that the effect on preventing death due to bleeding decreased with increasing time from injury (interaction p<0.0001). Early treatment ≤1 h from injury) significantly reduced the risk of deaths due to bleeding (relative risk 0.68), treatment given between 1 and 3 h also reduced the risk (RR 0.79), while treatment given after 3 h seemed to increase the risk of death due to bleeding (RR 1.44). The decrease across levels of the effect modifier strengthen the results.[52]

The effect modification persisted after adjustment for other potential effect modifiers: Credibility may be higher if a multivariable analysis suggests that the apparent effect modifier is independent of other candidate modifiers.[82] For example, a forward stepwise procedure may be used to investigate whether more than one modifier exists. Note that statistical independence of multiple effect modifiers does not guarantee a causal interpretation but makes it more likely. Most analysis of effect modification, however, do not have the power for meaningful multivariable analyses and the most relevant effect modifiers might be unknown.
Example: In a trial testing the administration of tranexamic acid in bleeding trauma patients found that the effect on preventing death due to bleeding decreased with increasing time from injury (interaction p<0.0001). Early treatment ≤1 h from injury) significantly reduced the risk of deaths due to bleeding (relative risk 0.68), treatment given between 1 and 3 h also reduced the risk (RR 0.79), while treatment given after 3 h seemed to increase the risk of death due to bleeding (RR 1.44). The investigators considered also three other potential effect modifiers. When including interaction terms for all four effect modifiers in a common model, the effect modification by time from injury remained highly significant (p<0.0001). [52]

Risk of bias of the main effect of the RCTs: We are less confident in any secondary analysis if studies are at high risk of bias with respect to random allocation, blinding,

missing data, and reporting. A commonly used instrument to formally assess the overall risk of bias is the Cochrane risk of bias tool.[83] There is, however, limited literature about the relationship between overall risk of bias and bias in analyses of effect modification. Some methodologists have argued that interaction tests are often robust to confounders of the main effect and measurement error of the effect modifier.[80] Some studies have suggested that industry funded trials are at higher risk of spurious claims of effect modification than non-industry funded studies, especially if the overall effect is not significant.[84-86]

The trial had had exceptionally high power to detect the effect modification: Methodologists have argued that the credibility of a proposed effect modification increases with its prospective power.[61,87] A rare situation of increased confidence would be a trial of over 10,000 patients with 80% power to detect a significant effect modification suggested in the study protocol.[61] Most analyses of effect modification, however, have low power and protocols rarely include an explicit power calculation.

The effect modification is consistent across related outcomes: Credibility might be higher if an effect modification is found for outcomes related biologically (or in another way). For instance, effect modifiers may be expected to have similar effects for stroke and myocardial infarction. Note that it is important to assess consistency by the size and direction of the effect modification and not by statistical significance alone which may be driven by differing sample sizes. Beware though that some biases may manifest across related outcomes and erroneously suggest increased credibility.
Example: In the trial of reamed versus unreamed nailing of tibial fractures, unreamed nailing apparently reduced re-operations in current smokers while reamed nailing reduced re-operations in other patients. The difference co-existed in other outcomes including quality of life measures Health Utility Index and short form-36; Results consistently suggested the superiority of unreamed nailing over reamed nailing in current smoking patients, and no or a small difference between unreamed and reamed nailing in other patients. This result strengthens the inference about the effect modification by smoking status.[34]

3.7 How would you rate the overall credibility of the proposed effect modification?

Item explanation: The instrument concludes with an overall credibility rating to summarize the considerations of the credibility questions.

The overall rating is a continuous visual analogue scale spanning four credibility areas. The credibility areas provide labels for credibility (the credibility areas roughly correspond to <25%, 25-50%, 50-75%, and >75% confidence that the apparent effect modification is true and not the result of chance or bias)

The overall rating should be driven by the items that decrease credibility. The following provides a sensible strategy:

- All responses definitely or probably reduced credibility or unclear → very low credibility
- Two or more responses definitely reduced credibility → maximum usually low credibility even if all other responses satisfy credibility criteria
- One response definitely reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- Two responses probably reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- No response options definitely or probably reduced credibility → high credibility very likely

It is helpful to justify the overall rating and weighting of items using the space provided below the overall rating.

Below the credibility labels, the scale provides an interpretation of the credibility rating (e.g. very low credibility suggests that there is very likely no effect modification) and implications for decision making (e.g. very low credibility implies that decision makers should consider the overall effect instead of subgroup-specific effects).

Section 5 provides more suggestions for using and presenting ICEMAN in context, section 6 more detailed justification why the scale is continuous and why low credibility suggests likely no effect modification.

Users can put a mark anywhere on the continuous line to rate the overall credibility (type “I” or “X” when using electronically).

It is helpful to justify the overall rating and weighting of items using the space provided below the overall rating.

3.8 Completed example for effect modification claimed in an RCT

A secondary publication of the CRASH-2 trial investigated the effect of tranexamic acid (an antifibrinolytic agent) versus placebo on death due to bleeding in trauma patients.[52] The investigators reported that "the effect of tranexamic acid on death due to bleeding varied according to the time from injury to treatment (test for interaction p<0·0001)." Although the investigators label their analysis as exploratory, an assessment using ICEMAN suggests moderate credibility.

Preliminary considerations

Study reference(s): Main publication: (Lancet 2010; 376: 23–32), CRASH 2 trial; secondary publication focussing on subgroup effect of interest: Lancet 2011; 377: 1096–101 "The importance of early treatment with tranexamic acid in bleeding trauma patients: an exploratory analysis of the CRASH-2 randomised controlled trial"

If available, protocol reference(s): https://www.thelancet.com/protocol-reviews/05PRT-1

State a single outcome and, if applicable, time-point of interest (e.g. mortality at 1 year follow-up): Death due to bleeding within 8 hours after injury

State a single effect measure of interest (e.g. relative risk or risk difference): Risk ratio

State a single potential effect modifier (e.g. age or comorbidity): Time from injury to treatment

Was the proposed effect modifier measured before randomization? [**X**] yes, continue [ ] no, stop here and refer to manual for further instructions

Credibility assessment

1: Was the direction of the effect modification correctly hypothesized a priori?

[ ] Definitely no	[ ] Probably no or unclear	[X] Probably yes	[ ] Definitely yes
Clearly post-hoc or results inconsistent with hypothesized direction or biologically very implausible	*Vague hypothesis or hypothesized direction unclear*	*No prior protocol available but unequivocal statement of a priori hypothesis with correct direction of effect modification*	*Prior protocol available and includes correct specification of direction of effect modification, e.g. based on a biologic rationale*

Comment: Subgroups pre-specified in published protocol, explicit statement that they had anticipated the direction of smaller effect on preventing death due to bleeding with increasing time after injury (although surprised by qualitative subgroup effect); hypothesis not stated in available protocol

2: Was the effect modification supported by prior evidence?

[ ] Inconsistent with prior evidence	[X] Little or no support or unclear	[ ] Some support	[ ] Strong support
Prior evidence suggested different direction of effect modification	*Consistent with weak or very indirect prior evidence (e.g. animal study at high risk of bias) or unclear*	*Consistent with more limited or indirect prior evidence (e.g. large observational study, non-significant effect modification in prior RCT, or different population)*	*Consistent with strong prior evidence directly applicable to the clinical scenario (e.g. significant effect modification in related RCT)*

Comment: The main paper cites three papers that seem to represent expert opinion but no prior cohort study or RCT

3: Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification? (consider irrespective of number of effect modifiers)

[ ] Chance a very likely explanation	[ ] Chance a likely explanation or unclear	[ ] Chance may not explain	[X] Chance an unlikely explanation
Interaction p-value > 0.05	*Interaction p-value ≤ 0.05 and > 0.01, or no test of interaction reported and not computable*	*Interaction p-value ≤ 0.01 and > 0.005*	*Interaction p-value ≤ 0.005*

Comment: Interaction p-value <0.00001

4: Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

[X] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Explicitly exploratory analysis or large number of effect modifiers tested (e.g. greater than 10) and multiplicity not considered in analysis	*No mention of number or 4-10 effect modifiers tested and number not considered in analysis*	*No protocol available but unequivocal statement of 3 or fewer effect modifiers tested*	*Protocol available and 3 or fewer effect modifiers tested or number considered in analysis*

Comment: Four pre-specified effect modifiers but applied to other outcomes than pre-specified in protocol (therefore labelled exploratory). The p-value is very small and conclusions unlikely to be altered if corrected for multiplicity. No original protocol available, only brief protocol on Lancet website

5: If the effect modifier is a continuous variable, were arbitrary cut points avoided? [ ] not applicable: not continuous

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[X] Definitely yes
Analysis based on exploratory cut point (e.g. picking cut point associated with highest interaction p-value)	*Analysis based on cut point(s) of unclear origin*	*Analysis based on pre-specified cut points, e.g. suggested by prior RCT*	*Analysis based on the full continuum, e.g. assuming a linear or logarithmic relationship*

Comment: Authors present continuous analysis of effect modification and a plot with 95% confidence bands

6 Optional: Are there any additional considerations that may increase or decrease credibility? (manual section 3.6)

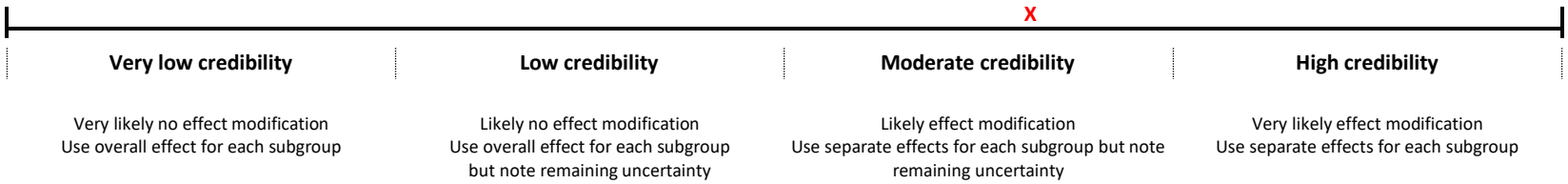
[ ] yes, probably decrease credibility [**X**] yes, probably increase credibility

Comment: "The evidence for interaction remained strong even after adjustment for interactions between the other pre-specified baseline characteristics and treatment (p<0.0001)"; apparent "dose-response" effect

7: How would you rate the overall credibility of the proposed effect modification?
The overall rating should be driven by the items that decrease credibility. The following provides a sensible strategy:

- All responses definitely or probably reduced credibility or unclear → very low credibility
- Two or more responses definitely reduced credibility → maximum usually low credibility even if all other responses satisfy credibility criteria
- One response definitely reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- Two responses probably reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- No response options definitely or probably reduced credibility → high credibility very likely

Put a mark on the continuous line (e.g. hit "x" in electronic version)



Comment: Lack of prior evidence and multiple testing are potential limitations but the p-value is very small and the continuous analysis reassuring

4 ICEMAN for meta-analyses

4.1 Is the effect modification based on comparison within rather than between RCTs?

Item explanation: Effect modification suggested by a comparison between studies (i.e. subgroups of studies) are usually much less credible than effect modification suggested by a comparison within studies (i.e. subgroups of individuals).

An important concern with between-study comparisons is study-level confounding: an association observed between a study level variable (e.g. type of intervention) and an outcome may be confounded by other study level variables (e.g. risk of bias).[8,68,88-95] If so, the apparent effect modification may be spurious. Study-level confounding might be particularly misleading when the putative effect modifier is a study-level summary of a participant-level variable (e.g. mean age or proportion of men). The study-level summary will often vary little across studies and will not reflect the true variation within studies. As a consequence, the power to identify a true within-trial effect modification can be very low and an apparent effect modification might be largely driven by study-level confounding.[94-96].

Most common are aggregate-data meta-analysis in which analyses of effect modification are usually completely based on between-study comparisons, e.g. using meta-regression. Those analysis are at a high risk of study-level confounding and consequently lower credibility.

Sometimes, investigators combine within and between-trial information using one of the following approaches:[68,97] 1) Estimate within- and between-trial effect modification separately, then combine both; 2) include a simple interaction term in a one-stage IPD meta-analysis; 3) first combine trials within subgroups, then compare summary effects between subgroups. The third approach is the most flexible because it allows inclusion of trials that provide information on one subgroup (at the cost of a higher risk of study-level confounding).[68,97]

All three combined approaches are at some risk of study level-confounding, which users of ICEMAN can judge using the two middle response options: If the most influential studies contribute data to one subgroup only, then the effect modification might be driven by between-study differences and the credibility is probably decreased (check mostly between). If the most influential studies provide within-trial information, then the effect modification is likely driven by within-study information and the credibility probably increased (check mostly within). This is the case for most individual patient data meta-analyses: A survey of published IPD meta-analyses suggested that only a small proportion of analyses of effect modification separate within- from between-trial information; instead, most analyses seem to combine within and between trial information.[68] Therefore, unless there is a statement to the contrary, analyses of effect modification in an IPD meta-analysis likely combine within and between trial information and might not be free of study-level confounding.

An analysis of effect modification is definitely free of study-level confounding if it is completely based on within-trial information. This is possible if all trials in a meta-analysis provide (or allow

estimation of) within-trial effect modification (e.g. a ratio of risk ratios) and, in a separate step, one combines the estimates across trials.[68,97,98] Alternatively, there are more complex methods available for individual-participant data meta-analyses to separate out within-trial effect modification in a one-stage model.[68,97,99]

Response options and examples:

[ ] Completely between. Subgroup analysis or meta-regression comparing overall effects of each individual trial. This is typical for aggregate data meta-analysis.
Example: In a meta-analysis assessing the effect of inpatient versus usual care, patients undergoing orthopedic focused rehabilitation had a substantially larger functional benefit than patients undergoing geriatric focused rehabilitation (interaction p = 0.01).[100]The analysis was based on between-study comparison only and therefore at high risk of confounding. The individual studies may differ in many other ways than type of rehabilitation (e.g. type of participants or type of usual care), especially considering the complexity of the intervention. The apparent effect modification may therefore not translate to the individual patient.
Example 2: An individual patient data meta-analysis based on three RCTs suggested that mobile phone text messages can improve the adherence to antiretroviral therapy. An analysis of effect modification suggested that interactive messaging (i.e. patients can interact with health care providers by responding to the text messages) is more effective than passive information only (interaction p-value 0.01). Because the type of text message varied only between but not within studies, the significant interaction reflects a between study comparison at high risk of study-level confounding. The example shows that use of individual participant data does not guarantee that an analysis of effect modification is based on within-trial information.

[ ] Mostly between or unclear*:* Subgroup analysis or meta-regression with most information coming from overall effects, but some trials providing within-trial subgroup information
Example 1: A meta-analysis assessing the effect of preoperative chemotherapy for gastroesophageal adenocarcinoma on survival combined individual patient and aggregate data.[101]. The analysis suggested a potentially larger treatment effect in tumors of the gastroesophageal junction than two other locations (interaction p=0.08). Two trials contributed within trial information to all three subgroups, three trials contributed within trial information to two subgroups, and five trials contributed data to one subgroup only. The investigators first combined trials within subgroups using a random effects model and then compared effects between subgroups – a method that explicitly combines within and between trial information. The apparent effect modification might be explained by study-level confounder, e.g. risk of bias.

[ ] Mostly within: Most trials providing within-trial subgroup information; or individual participant data analysis that combines within and between trial information

Example: A individual participant data meta-analysis combined 13 trials comparing radiochemotherapy versus radiotherapy alone in patients with cervical cancer. A subgroup analysis based on tumor stage (three ordered categories) suggested that the relative benefit of the combined therapy on survival decreased with increasing tumor stage. The authors first pooled subgroup specific effects of each trial, resulting in one pooled effect per subgroup, then applied a chi-square test for trend (p=0.017).[102] *This method combines within- and between trial information and is therefore potentially affected by study-level confounding*.[68]

[ ] Completely within: Individual participant data analysis that separates within from between trial information, e.g. meta-analysis of interactions
Example: A meta-analysis of individual patient data from 16 trials compared low intensity interventions for depression with usual care. The investigators found a significant effect modification by baseline severity, suggesting that patients who are more severely depressed at baseline demonstrate larger treatment effects than those who are less severely depressed. The investigators chose a model that estimated the effect modification within each trial and separated out between-trial comparisons. In addition, they included a forest plot illustrating the heterogeneity of effect modifications across trials.[103]

4.2 If two or more within-trial comparisons are available, is the effect modification similar from trial to trial?

Item explanation: Credibility of effect modification increases if the effect modification has been replicated across independent studies. Replication provides the strongest protection against random error and decreases the likelihood of confounding. Because replication is never perfect, the response options allow some graduation by considering the direction and magnitude of the observed effect modifications.

If the item applies, it is helpful to quantify the magnitude of effect modification for each trial, e.g. by calculating a ratio or risk ratios for each trial (e.g. risk ratio in subgroup A over risk ratio in subgroup B[98]).

Note that this credibility consideration is *different* from assessing consistency (or heterogeneity) of treatment effects across studies (e.g. expressed by the I^2-measure[104]).

Response options and examples:

[ ] Definitely not similar: Effect modification reported for two or more trials and clearly different directions
[still looking for good example]

[ ] Probably not similar or unclear: Effect modification not reported for individual trials or too imprecise to tell
Example: A individual participant data meta-analysis combined 13 trials comparing radiochemotherapy versus radiotherapy alone in patients with cervical cancer. A subgroup analysis based on tumor stage suggested that the relative benefit of the combined therapy on survival decreased with increasing tumor stage (chi-square test for trend p=0.017). The authors reported the effect modification only for the combined dataset, not for individual trials. It was therefore not possible to assess consistency across trials.[102]

[ ] Mostly similar: Effect modification reported for two or more trials, mostly similar in direction, but considerable differences in magnitude
Example: A meta-analysis of individual patient data from 16 trials compared low intensity interventions for depression with usual care. The investigators found a significant effect modification by baseline severity, suggesting that patients who are more severely depressed at baseline demonstrate larger treatment effects than those who are less severely depressed. The investigators chose a model that estimated the effect modification within each trial and separated out between-trial comparisons. In addition, they included a forest plot illustrating the heterogeneity of effect modifications across trials. Considering the point estimates of the effect modifications within the 16 trials, 12 suggested a direction consistent with the overall, 1 suggested no effect modification, and 3 were in the opposite direction but with wide confidence intervals.[103]

[ ] Definitely similar: Effect modification reported for two or more trials, similar in direction, only some differences in magnitude
Example 1: An IPD meta-analysis of using fixed-dose aspirin for primary prevention of cardiovascular events found a significant interaction with body weight. When the dose was low (<100mg), only patients at low body weight (<70kg) had a benefit. In the supplement, they provided a within-trial subgroup stratified by trial using the hazard ratio scale. Although the effect modification was not significant in some trials, all six trials showed the same direction (more effective in lighter patients) with ratios of hazard ratios ranging between 0.5 and 0.9.[105]

4.3 For between-RCT comparisons, is the number of studies large?

Item explanation: For analysis of effect modification based on between-study comparisons, the credibility increases with the number of studies (analogous to number of observations in a regression analysis). If the number of observations is small, the proposed effect modification may result from overfitting or confounding. A large number of studies also increases the power of the analysis and improves modelling of between-study dispersion in a random effects model (see item 5.7).[43,55,87,106]

Response options are defined by a minimum number of studies in the smallest subgroup or, for continuous meta-regression, a minimum total number of studies in total. This approximate guidance may need modification in specific situations: When an effect modifier includes more

than two ordered categories, it might be acceptable if one of the subgroups includes a small number of studies. In continuous meta-regression, in addition to the number of studies, users should additionally consider how the studies are distributed across levels of the effect modifier. For instance, if the total is 20 studies but 18 of them cluster at one end of the spectrum, the evidence is much weaker than if the studies were more evenly distributed across the spectrum.

Response options and examples:

[ ] Very small: 1 or 2 or in smallest subgroup; 5 or less studies in continuous meta-regression
Example: A meta-analysis comparing transcatheter versus surgical aortic valve replacement found a qualitative interaction: transcatheter was superior to surgical if applied transfemoral, but inferior if applied transapical (interaction p=0.01 using random effect model). The smallest subgroup included only two studies.[107]

[ ] Rather small or unclear: 3-4 in smallest subgroup; 6-10 studies in continuous meta-regression
Example: In a meta-analysis investigating the effect of low-intensity pulsed ultrasound on bone healing, the subgroup of 12 studies at high risk of bias suggested a large benefit of ultrasound whereas the subgroup of 3 studies at low risk of bias suggested no benefit (interaction p<0.001). The rather small number of 3 studies in the smallest subgroup is a possible limitation of the otherwise convincing effect modification.[108]

[ ] Rather large: 5-9 in smallest subgroup; 11 to 15 in continuous meta-regression
Example: In a meta-analysis assessing the effect of inpatient rehabilitation versus usual care, patients undergoing orthopedic rehabilitation had a substantially larger benefit within one year than patient undergoing geriatric rehabilitation, showing an interaction p = 0.01. The investigators conducted a subgroup analysis using random-effect meta-regression. Both subgroups included 6 studies per subgroup. The relatively high number of studies per subgroup reduces the risk of study-level confounding (i.e. another factor than type of rehabilitations explaining the differences between subgroups) although uncertainty remains, especially in the context of complex interventions and usual care as a comparator.[100]

[ ] Large: 10 or more in smallest subgroup; more than 15 in continuous meta-regression
Example: A study-level meta-analysis comparing interventions for preventing hospital readmission after discharge versus standard care performed a subgroup analysis by number of activities included in the intervention. The subgroup analysis suggested that only intervention with 5 or more activities were better than standard care but not intervention with 4 or less activities (interaction p=0.001). Because of the high heterogeneity regarding components of interventions and control groups between studies, the risk of confounding by other study characteristics seems relatively high. It is therefore reassuring that the small subgroup included 16 and the larger subgroup 26 studies.[109]

4.4 Was the direction of effect modification correctly hypothesized a priori?

Item explanation: Credibility is higher if investigators correctly anticipated the direction of the effect modification (e.g. that an intervention is more effective in younger than in older patients), lower if they failed to anticipate a direction, and lowest if they anticipated the opposite direction.

This item captures a number of credibility considerations:

Correct anticipation of an effect modification implies that the investigators had a specific hypothesis in mind – usually based on a biologic or other causal rationale, or sometimes based on external evidence. For instance, investigators may have anticipated a stronger relative effect in younger than in older patients because a disease may be too advanced in older patients for the intervention to be effective. If the data conforms to this hypothesis, the credibility is increased, otherwise decreased.

If the a priori specification of the effect modification hypothesis does not include a direction (e.g. by specifying the in the protocol that that the effect may vary by age but failure to say whether the effect is greater in the old versus the young or the other way round) this is weaker and probably not much better than having no prior hypothesis at all. In the Bayesian framework, the idea of a specific a priori hypothesis corresponds to using an informative rather than non-informative prior.[22,23]

In addition, the item captures that an explanation (e.g. a biological rationale) stated a priori is much more credible than a post hoc explanation. If post hoc, investigators had likely considered many possible explanations, thereby creating a multiplicity problem.[6,9,24-28]

Moreover, the item captures that hypotheses are most credible if verified in a prior, ideally in a time-stamped protocol or analysis plan. Note that a statements of pre-specification may not be reliable,[29] nor do they imply a specific hypothesis.

Note that the direction of an effect modification may depend on the type of effect measure if the effect modifier is also a prognostic factor (most are, e.g. age).

Because meta-analyses are retrospective, a potentially relevant caveat is that the investigators may already know the key trials and most promising effect modifiers when they plan the analysis; [88] if so, this item loses some of its value if it suggests increased credibility. For instance, a large trial may suggest an important effect modification. A subsequent individual patient data meta-analysis, in which the trial is influential, will likely show a similar effect modification. If investigators know the trial beforehand, correct anticipation of direction would essentially be data-driven. The item is more relevant if none of the key trials has tested the effect modifier of interest, and if the analysis of effect modification is completely based on a between-trial comparisons.

Response options and examples:

[ ] Definitely no: Clearly post-hoc or results inconsistent with hypothesized direction or biologically very implausible
Example: The ISIS-2 trial demonstrated that treatment with Aspirin can substantially reduce the number of vascular deaths in patients with suspected myocardial infarction.[31] A nonsense post-hoc subgroup analysis by birth sign suggested that the benefit occurred in all patients but those born under the sign of Gemini and Libra who did not appear to benefit from Aspirin.[31] This paper has become a classical example demonstrating that post-hoc subgroup analyses can easily mislead.

[ ] Probably no or unclear: Vague hypothesis or hypothesized direction unclear
Example: An IPD meta-analysis of using fixed-dose aspirin for primary prevention of cardiovascular events found a significant interaction with body weight. When the dose was low (<100mg), only patients at low body weight (<70kg) had a benefit. The paper does not clarify whether the effect modification was hypothesized a priori.[105]

[ ] Probably yes: No protocol available but unequivocal statement of a priori hypothesis with correct direction of effect modification
Example: An IPD meta-analysis combined three trials comparing high versus low positive end-expiratory pressure in ventilated patients with lung injury or ARDS. A subgroup analysis suggested that higher pressure was associated with longer survival in patients with but not in patients without ARDS (interaction p=0.02). The authors explicitly stated that they correctly anticipated the effect modification in their protocol which, however, was not published.[110]

[ ] Definitely yes: Prior protocol available and includes hypothesis with correct specification of direction of effect modification, e.g. based on biologic rationale
Example: A meta-analysis comparing transcatheter versus surgical aortic valve replacement found a qualitative interaction: transcatheter was superior to surgical if applied transfemoral, but inferior if applied transapikal. The investigators had anticipated this interaction with correct direction in a published protocol.[107]

4.5 Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification?

Item explanation: Credibility is higher if statistical test for interaction suggests that chance is an unlikely explanation for the apparent effect modification.[42-44] Credibility is lower if an interaction test suggests that an apparent effect modification is compatible with chance – or no test is available and impossible to compute.

For this item, consider the results of the interaction test (usually a p-value) as reported, irrespective of whether the p-value was adjusted for the number of analyses or not, or effect modifiers were analyzed jointly or one-by-one. We deal with considerations of multiple analyses separately in the following item.

Note that showing that an effect is significant in one subgroup and not in another is of little use: it provides no information whether chance might explain differences in effects across subgroups.[9,10,43,45,46]

There are a number of tests available including a chi square test, a chi square test of trend for ordered categories, or meta-regression for study-level analysis, or, if individual participant data is available, an interaction term in a one stage regression model, or a meta-analysis of trial-level interactions among other options.[97]

If no interaction p-value is reported, it can sometimes be calculated based on the reported data (point estimates of effect and confidence intervals in individual subgroups).[47,48] As rule of thumb is that the interaction p-value must be smaller than 0.05 if 95% confidence intervals of subgroup-specific estimates do not overlap.

We anchored the response options around typical thresholds for p-values 0.05, 0.01, and 0.005, with a p-value of 0.005 or smaller representing the most credible category. The response options recognize that p-value thresholds of 0.05 or even 0.01 may be too lenient for claiming statistical significance.[49] Of course, these are arbitrary settings and some methodologists would recommend even lower thresholds. Because of the low power of many analyses of effect modification, however, this would decrease the responsiveness of the item and the instrument's ability to distinguish more from less credible effect modification.

Note that other statistical measures than p-values such as interaction confidence intervals or Bayes factors may be more informative and intuitive than p-values but are rarely reported.

<u>Response options and examples:</u>

> **[ ] Chance a very likely explanation:** Interaction p-value > 0.05
> *Example: A meta-analysis assessed the effect of preoperative chemotherapy for gastroesophageal adenocarcinoma on survival explicitly. The investigators combined trials within subgroups according to tumor site using a random effects model, and then compared effects between subgroups using a chi-square test. The analysis suggested larger treatment effects in tumors of the gastroesophageal junction than other locations, but the p-value was only 0.08. Appropriately, the investigators emphasized that the finding requires prospective confirmation.*[101]
>
> **[ ] Chance a likely explanation or unclear:** Interaction or meta-regression p-value ≤ 0.05 and > 0.01, or no test of interaction reported and not computable
> *Example: An individual patient data meta-analysis investigated the effects of adding whole-brain radiation therapy to stereotactic surgery of brain metastases. An analysis of effect modification treating age as a continuous effect modifier suggested a lower mortality of surgery alone in younger patients, but the effect disappeared with increasing age. The p-value of 0.04 for the interaction term provided only little support against chance.*[111]

[ ] Chance may not explain: Interaction or meta-regression p-value ≤ 0.01 and > 0.005
Example: In a meta-analysis assessing the effect of inpatient rehabilitation versus usual care, patients undergoing orthopedic-focused rehabilitation had a substantially larger improvement in function 3-12 months after randomization than patient undergoing geriatric-focused rehabilitation. A random effects meta-regression analysis showed an interaction p-value of 0.01.[100]

[ ] Chance an unlikely explanation: Interaction or meta-regression p-value ≤ 0.005
Example: An individual participant data meta-analysis combining trials comparing low-dose aspirin versus placebo reported a subgroup analysis by body-weight. The interaction test suggested that aspirin reduced cardiovascular events in patients weighing less than 70kg but not in other patients. The interaction p-value of 0.007 suggested substantial support against chance.[105]

4.6 Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

Item explanation: Performing multiple tests is a major concern in the context of effect modification analysis. Trialists usually measure a large number of baseline variables, many of which they could test for potential effect modification. Because multiple tests increase the risk of a chance finding,[53-55] credibility is higher if investigators have tested only a small number of effect modifiers. Conversely, credibility decreases with the number of tested candidate effect modifiers. We therefore advise counting the number of candidate effect modifiers stated, ideally verified in a protocol.

Multiplicity issues can arise in different ways.[56] Most obvious are situations in which investigators test multiple candidate effect modifiers and highlight significant results. Another important issue which we address in a separate item concerns selection of cut points of continuous effect modifiers. Other potential multiplicity issues include multiple time points, multiple scales,[57] multiple outcomes, or multiple methods for testing the interaction. Therefore, even if the number of effect modifiers is small, one should consider whether other issues might have introduced multiplicity.

An alternative to limiting the number of analyses is to statistically adjust the analysis for multiplicity. Credibility is higher if an effect modification persists after adjustment. Different techniques are available including correction of p-values considering the (familywise) type 1 error rate,[58] testing all candidate effect modifiers in a common model, using a composite variable such as a risk score, or shrinkage estimators.[45,59] All techniques inevitably reduce power.[6,60,61] Most, investigators, however, do not address potential multiplicity issues in design or analysis and leave the judgement to the reader – another reason why a small number of effect modifiers is most helpful.

Assessment of multiplicity crucially depends on reporting (reporting guidelines for effect modification are available [62-64]). Without knowing the number of effect modification analyses

performed, we cannot assess the potential impact of multiplicity. Ideally, investigators would specify candidate effect modifiers along with definitions and analytic details in a protocol. If no protocol is available, one should look for explicit statements about the number of effect modifiers. A note of caution: an empirical study has shown that retrospective statements about the number of pre-specified subgroup analyses are not always reliable.[29] Also note that a statement that a particular effect modifier was pre-specified does not rule out the problem of multiplicity because investigators may have pre-specified many other effect modifiers.

In summary, this item requires counting the number of effect modifiers (perhaps considering additional multiplicity issues), if possible verifying them in a protocol, and considering whether investigators considered the number of analyses in their statistical analysis.

A potential limitation is that the meta-analysts might have scanned key trials for promising effect modifiers before they planned the meta-analysis. If so, a small number of tested effect modifiers in a meta-analysis might obscure potential multiplicity issues introduced in earlier selection processes in the individual trials.

Response options and examples:

[ ] Definitely no: Explicitly exploratory analysis or large number of effect modifiers tested (e.g. greater than 10) and multiplicity not considered in analysis
Example: A meta-analysis investigating interventions to reduce early hospital readmissions reported results for 12 effect modifiers. One analysis suggested that interventions with at 5 or more components were more effective than interventions with less than 5 components (interaction p=0.001). The authors correctly highlighted the possibility of a chance findings due to multiplicity. Sources of multiplicity were the number of tested effect modifiers and, for this particular effect modifier, choice of cut point.[109]

[ ] Probably no or unclear: No mention of number or 4-10 effect modifiers tested and number not considered in analysis
Example: In a meta-analysis assessing the effect of inpatient rehabilitation versus usual care, patients undergoing orthopedic rehabilitation had a substantially larger improvement in function that patient undergoing geriatric rehabilitation. A random effects meta-regression analysis showed an interaction p-value of 0.01. According to the authors, all reported meta-regression analyses were pre-specified in an analysis plan, increasing the confidence that selective reporting is unlikely.[100] Nevertheless, they tested 9 effect modifiers for 3 outcomes at 2 time points, most of which were not significant. The multiple analyses increase the risk of finding a spurious result as extreme as p=0.01.

[ ] Probably yes: No protocol available but unequivocal statement of 3 or fewer effect modifiers tested
[still looking for good example

[ ] Definitely yes: Protocol available and 3 or fewer effect modifiers tested or number considered in analysis *Example: A meta-analysis comparing the effect of low-intensity pulsed ultrasound versus sham ultrasound on bone healing reported a convincing effect modification: studies at high risk of bias suggested a large while studies at low risk of bias no treatment effect (p<0.001 using random-effects meta-regression). The investigators had pre-specified the analysis in the published protocol together with two other subgroup hypotheses. In addition, the protocol provided explicit criteria for classifying trials into high or low risk of bias.[112] The low number of tested effect modifiers and the pre-specified definition makes multiplicity issues less likely.[108]*

4.7 Did the authors use a random effects model?

Item explanation: The credibility of claimed effect modification increases if investigators used a random effects model for between-study differences that allows true effects to differ among studies. The credibility is lower if investigators used a fixed (or common) effect, or fixed effects (plural) model within subgroups; those models do not appropriately address uncertainty due to heterogeneity between studies.[113] More appropriate is a random-effects (or, more precisely, mixed effects) model.[113]) Simulation studies have shown that failure to assume random effects increases the risk of bias and false positive claims for both for study-level and individual-patient data meta-analysis because standard errors are underestimated. [55,99,106] A random effect model strengthens a test of interaction because a significant result is harder to achieve.[55,88,91,113,114]

If investigators just state that they used a random-effects model without further specification, it is reasonable to assume that they refer to a random-effects model for between-study differences within subgroups. When investigators do not sufficiently describe the methods, one may deduce that they probably employed a random-effects model, e.g. when they specified a random-effects model in their protocol for analyzing the main effect.

If investigators state that they used a mixed effects model without further specification, it usually implies that they used a random effects model for between-study differences and a fixed effects model for between-subgroup differences (the latter being appropriate.[91,113]) Therefore, the appropriate answer is usually definitely yes.

The question also applies to individual-participant data meta-analysis. An empirical study has shown that most IPD meta-analyses do not apply a random effects model.[115]

Response options and examples:

[ ] Definitely no: Fixed (or common) effect(s) is explicitly stated.
Example: Example: An IPD meta-analysis of using fixed-dose aspirin for primary prevention of cardiovascular events found a significant interaction with body weight and age. The authors explicitly used a fixed effects model.[105]

[ ] Probably no or unclear: Probably fixed effects model or unclear.

Example: A individual participant data meta-analysis combined 13 trials comparing radiochemotherapy versus radiotherapy alone in patients with cervical cancer. A subgroup analysis based on tumor stage suggested that the relative benefit of the combined therapy on survival decreased with increasing tumor stage. The authors used the chi-square test for trend (p=0.017) but did not explicitly report whether the analysis of effect modification was based on a fixed effects model, as in their primary analysis, or random effect model as in their secondary analysis. Most likely, they used a fixed effect model as for the primary analysis.[102]

[ ] Probably yes: Probably random (or mixed) effects model.
[still looking for good example]

[ ] Definitely yes: Random (or mixed) effects model is explicitly stated.
Example: In a meta-analysis assessing the effect of inpatient rehabilitation versus usual care, patients undergoing orthopedic rehabilitation had a substantially larger improvement sin function that patient undergoing geriatric rehabilitation. A meta-regression analysis showed an interaction p-value of 0.01. In the methods, the authors explicitly specify a random effects model for between study differences.

4.8 If the effect modifier is a continuous variable, were arbitrary cut points avoided?
Item explanation: Categorizing continuous effect modifiers is common[68] but associated with a number of problems:[69,70] Cut points can introduce multiplicity, reduce power, mask linear or non-linear associations. In the context of meta-analysis, cut points can cause additional problems. If two studies assessed the same continuous effect modifier but used different cut points, it may be impossible to combine the (within-study) results in a meaningful way unless individual patient-data is available. Therefore, analyses that avoid cut points and make use of the full spectrum of values are the most credible.

Investigators often decide against using the complete data and rather use cut points to partition continuous effect modifiers in two or more categories. Categories with a strong, empirically grounded rationale, are the most credible. For instance, arbitrariness can be avoided by pre-specifying the cut points based on a previous RCT that demonstrates effect modification. Credibility is low if investigators selected the best-fitting data-driven cut point to maximize the effect modification. Such cut points are associated with a high rate of false positive claims.[69,71]

There are some challenges when modelling continuous effect modifiers that are not part of the instrument but may lower the credibility: model misspecification can occur if the continuous relationship is driven by a few influential observations.[72-74] Post-hoc modelling can lead to overfitting. Most credible are therefore continuous analyses for which investigators have pre-specified the type of dependency of the treatment effect on the continuous variable (sometimes referred to as *treatment effect function*) such as a linear or log relationship, or considered a small number of candidate functions.[75]

An alternative to use of cut points and potentially complex modelling is to consider overlapping subgroups (e.g. using a sliding window approach).[76] The credibility is usually much higher than using arbitrary cut points but the interpretation can be difficult.

The credibility of a continuous analysis usually increases if investigators present a plot with confidence bands around the regression function (often a line) and carefully checked the proposed model. Provided individual participant data is available, it is also possible to average functions across several studies and base conclusions on the resulting mean function (i.e. a meta-analysis of interactions, see item 4.1).[116,117] Credibility increases if most of the individual function show a similar relationship between the continuous variable and the outcome (see item 4.2).

Note that additional considerations related to continuous effect modifiers may apply, e.g. if there is a clear dose-response relationship or results were robust to sensitivity analyses (see following question).

Response options and examples:

[ ] Definitely no: Analysis based on exploratory cut point(s), e.g. picking cut point associated with highest interaction p-value
[still looking for good example]

[ ] Probably no: Analysis based on cut point(s) of unclear origin
Example: A meta-analysis investigating interventions to reduce early hospital readmissions reported a potential effect modification by the number of intervention components. Studies with Interventions with 5 or more components showed a significant effect while studies with less than 5 components showed no significant effect (interaction p=0.001). The published protocol did not specify cut points and the investigators explicitly highlighted the exploratory character of the analysis. Presentation of different cut points or treating the effect modifier as a continuous variable would have been reassuring. [109]

[ ] Probably yes: Analysis based on pre-specified cut point(s), e.g. suggested by prior RCT
Example: In a meta-analysis on inpatient rehabilitation versus usual care in elderly patients, the intervention was better in preventing nursing home admissions in patients younger than 80 than in patients older than 80 (p=0.045).[100] According to the authors, the threshold was pre-specified thus avoiding multiplicity due to arbitrary selection of cut points. There is some uncertainty as no protocol is available.

[ ] Definitely yes: Analysis based on the full continuum, e.g. assuming a linear or logarithmic relationship
Example: An IPD meta-analysis investigated whether patients with acute respiratory distress syndrome (ARDS) benefit from higher positive end-expiratory pressure (PEEP)

ventilation strategies. A continuous analysis of effect modification suggested a nonlinear effect modification by degree of hypoxaemia (expressed as the ratio of PaO_2/FiO_2. A higher PEEP reduced mortality only in patients with values between 100 and 150 but not in patients with lower values.[117] A previous analysis dichotomized the effect modifier and could not reveal the potential non-linear relationship.[110] The investigators also provided plots of the proposed effect modification including confidence limits (suggesting high uncertainty in this case).[117]

4.9 Optional: Are there any additional considerations that may increase or decrease credibility?

Item explanation: Methodologists have suggested a number of additional considerations that could be relevant for assessing the credibility of effect modifiers.[118] They are not part of the core items because they either are less relevant, rarely apply, or are difficult to assess. Because they are usually less relevant than the previous core items, the only response options are probably decreased and probably increased.

Additional considerations are optional, that is, leaving this section blank does not affect credibility. Note that it may not be worth to consider potential additional considerations if core items already suggest low or very low credibility.

The following list provides potentially relevant additional considerations:

A sensitivity analysis suggested robustness to relevant assumptions: A sensitivity analysis can help to increase the confidence in a proposed effect modification.[15,80,81] For instance, if an effect modification analysis is based on a dichotomized continuous variable, the credibility increases if the effect modification persists for different cut-points.
Example: A meta-analysis comparing the effect of low-intensity pulsed ultrasound versus sham on bone healing reported a convincing subgroup effect: studies at high risk of bias suggested a large while studies at low risk of bias consistently suggested no effect (interaction p<0.001 based on univariable random-effects meta-regression). Part of the criteria for classifying trials into high and low risk of bias was 20% or more missing data. In a sensitivity analysis requested by the editors, the investigators applied a more strict threshold for missing data (≥10%). Although the different criteria led to reclassification of one trial from low to high risk of bias, the effect modification remained significant (p=0.004). The sensitivity analysis increased the confidence of the editors that the effect modification was real.[108]

Effect modification supported by external evidence: The credibility may be higher if the proposed effect modification is consistent with findings from studies that are not included in the meta-analysis, e.g. a high quality cohort study.
Example: A meta-analysis comparing transcatheter versus surgical aortic valve replacement found a qualitative interaction: transcatheter was superior to surgical if applied transfemoral, but inferior if applied transapical. A prior cohort study of 501

patients (i.e. data not included in the meta-analysis or RCTs) using propensity score matching had suggested that the transapical approach was associated with more adverse events and higher mortality than the transfemoral approach.[119]

"Dose-response effect" across levels of the effect modifier: Credibility may be higher if effects increase or decrease monotonically with increases in the levels of the modifier, e.g. an effect that increases incrementally across three or more age groups. On the contrary, it is especially important to beware of apparent effect modification that do not reflect a plausible pattern across three or more ordered groups, even if statistically significant. For instance, an effect might me abnormally elevated in one subgroup chosen from a continuum, but not in neighboring subgroups.
Example: A individual participant data meta-analysis combined 13 trials comparing radiochemotherapy versus radiotherapy alone in patients with cervical cancer. A subgroup analysis based on tumor stage suggested that the relative benefit of the combined therapy on survival decreased with increasing tumor stage (across three stages), suggesting a possible "dose-response" effect (chi-square test for trend, p=0.017).[102]

Risk of bias of the main effects of the individual RCTs or the meta-analysis: We are less confident in an analysis of effect modification if the individual studies or the meta-analysis itself is at high risk of bias. A commonly used instrument to formally assess the overall risk of bias is the Cochrane risk of bias tool for individual trials[83] and the ROBIS tool for systematic reviews.[120] There is, however, limited literature about the relationship between overall risk of bias and bias in analyses of effect modification. Some methodologists have argued that interaction tests are often robust to confounders of the main effect and measurement error of the effect modifier.[80] Note that reporting bias can be introduced if only some studies report an effect modifier but not others as reporting is likely driven by the results.[121] Also, meta research has suggested that industry funded trials are at higher risk of spurious claims of effect modification than non-industry funded studies, especially if the overall effect is not significant.[84-86]
Example: An IPD meta-analysis combined three trials comparing high versus low positive end-expiratory pressure in ventilated patients with lung injury or ARDS. A subgroup analysis suggested that higher pressure was associated with longer survival in patients with but not in patients without ARDS (interaction p=0.02). Although the p-value provides only modest support against chance, the high methodological quality of all three trials is reassuring.[110]

The meta-analysis had had exceptionally high power to detect the effect modification: Methodologists have argued that the credibility of a proposed effect modification increases with its prospective power.[61,87] A rare situation of increased confidence would be an IPD meta-analysis of over 10,000 patients with 80% power to detect a significant effect modification suggested in the study protocol.[61] Most analyses of effect

modification, however, have low power and protocols rarely include an explicit power calculation.

The effect modification persisted after adjustment for other potential effect modifiers: Credibility may be higher if a multivariable analysis suggests that the apparent effect modifier is independent of other candidate modifiers.[82] Note that statistical independence of multiple effect modifiers does not guarantee a causal interpretation but makes it more likely. Most analysis of effect modification, however, do not have the power for meaningful multivariable analyses and the most relevant effect modifiers might be unknown.
Example 1: An IPD meta-analysis of using fixed-dose aspirin for primary prevention of cardiovascular events found a significant interaction with body weight and age. The effect modification by weight remained when the investigators stratified their analysis by both variables.[105]

The effect modification is consistent across related outcomes: Credibility might be higher if an effect modification is found for biologically (or in another way) related outcomes. For instance, effect modifiers may be expected to have similar effects for stroke and myocardial infarction. Note that it is important to assess consistency by the size and direction of the effect modification and not by statistical significance alone which may be driven by differing sample sizes. Beware though that some biases may manifest across related outcomes and erroneously suggest increased credibility.
Example: A meta-analysis comparing transcatheter versus surgical aortic valve replacement found a qualitative interaction: transcatheter was superior to surgical if applied transfemoral, but inferior if applied transapical. The interaction was consistent across outcomes mortality, stroke, acute kidney injury, and bleeding.[107]

4.10 How would you rate the overall credibility of the proposed effect modification?

Item explanation: The instrument concludes with an overall credibility rating to summarize the considerations of the credibility questions.

The overall rating is a continuous visual analogue scale spanning four credibility areas. The credibility areas provide labels for credibility (the credibility areas roughly correspond to <25%, 25-50%, 50-75%, and >75% confidence that the apparent effect modification is true and not the result of chance or bias)

The overall rating should be driven by the items that decrease credibility. The following provides a sensible strategy:

- All responses definitely or probably reduced credibility or unclear → very low credibility
- Two or more responses definitely reduced credibility → maximum usually low credibility even if all other responses satisfy credibility criteria
- One response definitely reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- Two responses probably reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- No response options definitely or probably reduced credibility → high credibility very likely

It is helpful to justify the overall rating and weighting of items using the space provided below the overall rating.

Below the credibility labels, the scale provides an interpretation of the credibility rating (e.g. very low credibility suggests that there is very likely no effect modification) and implications for decision making (e.g. very low credibility implies that decision makers should consider the overall effect instead of subgroup-specific effects).

Section 5 provides more suggestions for using and presenting ICEMAN in context, section 6 more detailed justification why the scale is continuous and why low credibility suggests likely no effect modification.

Users can put a mark anywhere on the continuous line to rate the overall credibility (type “I” or “X” when using electronically).

It is helpful to justify the overall rating and weighting of items using the space provided below the overall rating.

4.11 Example for an effect modification claimed in a meta-analysis

An individual patient data meta-analysis of 13 trials compared radiochemotherapy versus radiotherapy alone in women with cervical cancer. The authors report in their abstract “a suggestion of a difference in the size of the survival benefit with tumor stage”. The credibility assessment suggested low credibility for the proposed effect modification.

Preliminary considerations

Study reference(s): J Clin Oncol 2008 26:5802-5812, radiochemotherapy versus radiotherapy alone in women with cervical cancer

If available, protocol reference(s): Not published, publication states available on request

State a single outcome and time-point of interest: Death

State a single effect measure of interest (e.g. relative risk or risk difference): Hazard ratio

State a single proposed effect modifier (e.g. age or comorbidity): Tumor stage (three ordered stages)

Was the effect modifier measured before randomization? [**X**] yes, continue [ ] no, stop here and refer to manual for further instructions

1: Is the analysis of effect modification based on comparison within rather than between trials?

[ ] Completely between	[ ] Mostly between or unclear	[X] Mostly within	[ ] Completely within
Subgroup analysis or meta-regression comparing overall effects of each individual trial. This is typical for aggregate data meta-analysis.	*Subgroup analysis or meta-regression with most information coming from overall effects, but some trials providing within-trial subgroup information*	*Most trials providing within-trial subgroup information; or individual participant data analysis that combines within and between trial information*	*Individual participant data analysis that separates within from between trial information, e.g. meta-analysis of interactions*

Comment: All trials provided individual participant data. The authors probably first pooled trials within subgroups, then compared pooled effects between subgroups. This method combines within and between study information, although the suggested effect modification is likely driven by within-study information

2: For within-trial comparisons, is the effect modification similar from trial to trial? [ ] Not applicable: no or one within-RCT comparison

[ ] Definitely not similar	[X] Probably not similar or unclear	[ ] Mostly similar	[ ] Definitely similar
Effect modification reported for two or more trials and clearly different directions	*Effect modification not reported for individual trials or too imprecise to tell*	*Effect modification reported for two or more trials, mostly similar in direction, but considerable differences in magnitude*	*Effect modification reported for two or more trials, similar in direction, only some differences in magnitude*

Comment: No information

3: For between-trial comparisons, is the number of trials large? [ ] Not applicable: no between RCT comparison

[ ] Very small	[ ] Rather small or unclear	[X] Rather large	[ ] Large
1 or 2 or in smallest subgroup; 5 or less in continuous meta-regression	*3-4 in smallest subgroup; 6-10 in continuous meta-regression*	*5-9 in smallest subgroup; 11 to 15 in continuous meta-regression*	*10 or more in smallest subgroup; more than 15 in continuous meta-regression*

Comment: 13 trials

4: Was the direction of effect modification correctly hypothesized a priori?

[ ] Definitely no	[X] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Clearly post-hoc or results inconsistent with hypothesized direction or biologically very implausible	*Vague hypothesis or hypothesized direction unclear*	*No prior protocol available but unequivocal statement of a priori hypothesis with correct direction of effect modification*	*Prior protocol available and includes correct specification of direction of effect modification, e.g. based on a biologic rationale*

Comment: No information

5: Does a test for interaction suggest that chance is an unlikely explanation of the apparent effect modification? (consider irrespective of number of effect modifiers)

[ ] Chance a very likely explanation	[X] Chance a likely explanation or unclear	[ ] Chance may not explain	[ ] Chance an unlikely explanation
Interaction or meta-regression p-value > 0.05	*Interaction or meta-regression p-value ≤ 0.05 and > 0.01, or no test of interaction reported and not computable*	*Interaction or meta-regression p-value ≤ 0.01 and > 0.005*	*Interaction or meta-regression p-value ≤ 0.005*

Comment: P=0.017 for chi-square test of trend

6: Did the authors test only a small number of effect modifiers or consider the number in their statistical analysis?

[ ] Definitely no	[X] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Explicitly exploratory analysis or large number of effect modifiers tested (e.g. greater than 10) and multiplicity not considered in analysis	*No mention of number or 4-10 effect modifiers tested and number not considered in analysis*	*No protocol available but unequivocal statement of 3 or fewer effect modifiers tested*	*Protocol available and 3 or fewer effect modifiers tested or number considered in analysis*

Comment: At least 8 subgroup analyses; no published protocol

7: Did the authors use a random effects model?

[ ] Definitely no	[**X**] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Fixed (or common) effect(s) explicitly stated	*Probably fixed (or common) effect(s)*	*Probably random (or mixed) effects*	*Random (or mixed) effects explicitly stated*

Comment: Not explicitly stated but authors used a fixed effects model for the overall analysis

8: If the effect modifier is a continuous variable, were arbitrary cut points avoided? [**X**] not applicable: not continuous

[ ] Definitely no	[ ] Probably no or unclear	[ ] Probably yes	[ ] Definitely yes
Analysis based on exploratory cut point(s), e.g. picking cut point associated with highest interaction p-value	*Analysis based on cut point(s) of unclear origin*	*Analysis based on pre-specified cut point(s), e.g. suggested by prior RCT*	*Analysis based on the full continuum, e.g. assuming a linear or logarithmic relationship*

Comment:

9 Optional: Are there any additional considerations that may increase or decrease credibility? (manual section 4.9)

[ ] yes, probably decrease credibility [**X**] yes, probably increase credibility

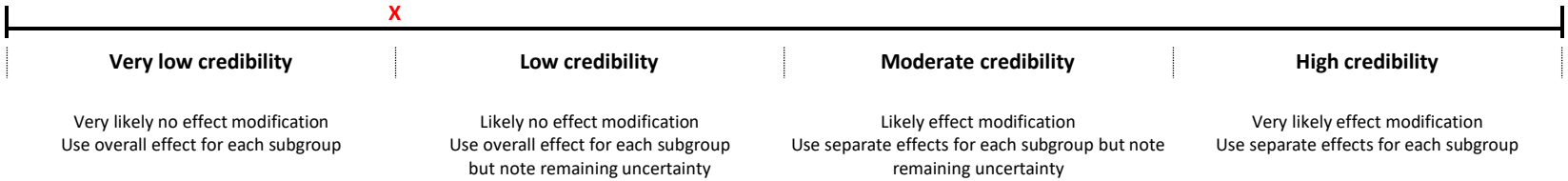
Comment: Possible "dose-response effect"; effect modification consistent across outcomes

10: How would you rate the overall credibility of the proposed effect modification?

The overall rating should be driven by the items that decrease credibility. The following provides a sensible strategy:

- All responses definitely or probably reduced credibility or unclear → very low credibility
- Two or more responses definitely reduced credibility → maximum usually low credibility even if all other responses satisfy credibility criteria
- One response definitely reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- Two responses probably reduced credibility → maximum usually moderate credibility even if all other responses satisfy credibility criteria
- No response options definitely or probably reduced credibility → high credibility very likely

Put a mark on the continuous line (e.g. hit "x" in electronic version)



Comment: Prior knowledge or evidence unclear; p-value not very small and possibly affected by multiple analyses; fixed effect model not optimal

5 Practical considerations

5.1 Assessment in duplicate

Confidence in the assessment increases if two investigators independently apply ICEMAN, discuss discrepancies, and present a version based on their consensus.

5.2 Reporting

We recommend specifying use of the instrument in the study protocol and in the methods, results, and interpretation sections of the final publication as in the following examples:

Study protocol: "We will assess the credibility of potentially relevant effect modification using ICEMAN.[citation]"

Methods section of publication: "We used ICEMAN[citation] to assess the credibility of potentially relevant effect modification."

Results section: "An analysis of effect modification treating age as a continuous effect modifier suggested that the benefit of the intervention diminished with increasing age of participants (Figure). We judged the credibility of the potential effect modification as low with uncertainty arising from lack of prior evidence and an inconclusive test of interaction (see supplement for detailed credibility assessment)."

Interpretation: "An analysis of effect modification suggested that the effect of the intervention might vary by age, but a formal credibility assessment rated the apparent effect modification as likely spurious. Therefore, we recommend considering the overall effect estimate for all patients, independent of their age."

When presenting the results of ICEMAN, we suggest sticking closely to the wording used in the instrument, which we developed based on user-testing.

We do *not* recommend reporting overall credibility as a percentage (e.g. *30% credible*).

5.3 Using ICEMAN in combination with other instruments

ICEMAN can be used in combination with an instrument to assess the risk of bias of main effects such as the Cochrane risk of bias tool for RCTs[83] or the ROBIS tool for meta-analyses.[120] If the overall risk of bias is low, use of ICEMAN is straightforward. If the overall risk of bias is substantial, there are three possible responses:

> 1) Do not apply ICEMAN because a rating of moderate or high credibility is unlikely, and evidence users are probably not interested in analyses of effect modification if the overall effect is uncertain.

2) Apply ICEMAN but mention the overall risk of bias as an additional source of uncertainty under *additional considerations*.

3) In the context of a meta-analysis, consider the individual studies' risk of bias as a potential effect modifier, perform a subgroup analysis based on risk of bias categories, and apply ICEMAN to assess credibility.

ICEMAN is compatible with the GRADE framework to rate the certainty of evidence and strength of recommendations as follows[122]:

1) ICEMAN suggests moderate or high credibility: Apply GRADE to subgroup-specific effects estimates. If moderate credibility, note remaining uncertainty. Considering subgroup-specific estimates may sometimes resolve concerns due to heterogeneity and consequently increase certainty of evidence and strength of recommendation. If the candidate effect modifier is methodological quality (e.g. risk of bias assessed by the Cochrane risk of bias tool), apply GRADE to the high-quality subgroup only.

2) ICEMAN suggests low or very low credibility: Apply GRADE to the overall effect estimate. If low credibility, note remaining uncertainty, especially if the potential effect modification appears to explain heterogeneity.

6 Additional conceptual considerations

The assessment assumes skepticism regarding possible effect modification: The instrument reflects the generally skeptical view on effect modification found in the theoretical literature and supported by meta-research, including the very small proportion of subgroup explorations that show apparent effect modification. Moreover, attempts to replicate subgroup effects are rare and, if undertaken, rarely successful.[36]

Several elements of the instrument reflect the general skepticism (equivalent to a skeptical prior in Bayesian terminology): the combination of response options unclear and probably decreased credibility; using relatively strict criteria for response options suggesting increased credibility (though some co-authors would have been even stricter); advice to base the overall credibility rating on response options suggesting decreased credibility (rather than averaging across individual items); and suggestions for interpreting low credibility as likely to indicate an absence of effect modification.

The assessment is about an association not a causal relationship: Effect modification refers to an association, not necessarily a causal relationship. For instance, a treatment effect may credibly vary among levels of a risk score, or body weight, although both are not causes of the effect modification. There might be other causal factors associated with both the apparent effect modifier and the outcome.[5,42,123,124] Unless the patients were randomized to subgroups

defined by the effect modifier, an analysis of effect modification resembles an observational study, even if applied within a randomized controlled trial.[5,123]
A causal interpretation becomes more likely if the ICEMAN rating is high credibility, but may nevertheless remain unlikely. The uncertainty regarding causality might have implications for further decision making, in particular if the putative effect modifier is an intervention characteristic. Causality is less critical if the effect modifier is a patient characteristic and the aim of the analysis is the identification of optimal patient subgroups for applying an intervention.[5]

Magnitude and relevance of effect modification are not part of the assessment: ICEMAN does not directly address the magnitude of effect modification. Therefore, it is usually not necessary to quantify the effect modification numerically, e.g. by specifying a ratio of risk ratios, or the value of an interaction term in a regression model. The only exception is the second item in the meta-analysis version that addresses consistency of effect modification across individual studies.

ICEMAN does not address whether a credible effect modification is important to the patient. Importance should be considered independently from credibility and depends on absolute effects, additional outcomes, and context.

It may be useful to consider importance of the potential effect modification to any potential course of action first before applying ICEMAN. If it is clear that consequences for decision making would not depend on the (potentially credible) effect modification, then it may not be worth investing in a credibility assessment. For instance, if an intervention compared to placebo shows a large effect in men and a very large effect in women, it might be unimportant to consider whether sex might be a credible effect modifier.

Choice of effect measure does not inform credibility: ICEMAN does not address whether a chosen effect measure (e.g. relative or absolute risk difference) is more or less appropriate. Credibility can be assessed on any scale of interest. The instrument addresses credibility on a particular scale that can be specified in the preliminary considerations.

There is no general consensus in the methodological literature on how to select the optimal effect measure.[125,126] One approach is – for binary outcomes – to generally prefer relative over absolute scales. Relative effects are more likely to be similar across baseline risk,[3,127] and as a result the heterogeneity of treatment effects is usually substantially lower if one chooses relative rather than absolute effects. The impact on heterogeneity is less clear for continuous outcomes.[3] Other authors generally prefer absolute effect measures such as risk differences,[126] which have some advantages (e.g. calculation of number needed to treat) but also disadvantages (e.g. higher heterogeneity across baseline risks makes it more difficult to summarize treatment effects as a single number and complicates meta-analysis).[127] A common recommendation is to analyze the data on a relative scale in which true effect modification is unusual, and then, for addressing the magnitude of effect in subgroups when effect

modification is credible, calculate magnitude of effects in each subgroup using an absolute scale.[59]

On using categorical and continuous rather than binary response options for addressing credibility of an effect modification claim: Discussions regarding the credibility of effect modification have often used polarizing terminology such as true positive versus false positive; confirmatory versus exploratory; or pre-specified versus ad hoc. In reality, however, any reasonable assessment of credibility will fall somewhere between definitely true and definitely false. Thus, a continuous, probabilistic concept is much more appropriate. ICEMAN uses four categorical response options for the core items and a continuous scale for the overall assessment divided into four areas. Making the overall assessment continuous instead of categorical results in higher formal ratings of reliability: when two raters differ on a four-point scale, they may in fact almost agree on a continuous scale. ICEMAN's four credibility areas facilitate reporting and are likely to be useful for consumers of the instrument ratings.

On the decision to offer two separate version for RCTs and meta-analyses of RCTs: In developing ICEMAN, we considered three main types of studies: individual RCTs, aggregate data meta-analyses, and individual participant data meta-analyses. We started with a version that combined all three types of studies but the complexity proved daunting. We also considered combining RCTs with individual participant data meta-analyses because both are based on individual participant data. Our final decision to separate individual trials and meta-analyses but not individual and aggregated data was mainly driven by the following considerations: 1) RCTs are prospective, meta-analyses are retrospective; this has consequences for the relative impact of a priori considerations and the concept of confirmation. 2) Most users are familiar with distinguishing individual trials from meta-analyses but many users are less familiar with the conceptual similarity of RCTs and individual participant data meta-analyses in the context of effect modification. 3) Individual participant and aggregate data meta-analysis is not mutually exclusive and combinations of both are possible.

7 References

1.Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3 ed. Philadelphia: Wolters Kluwer Health; 2012.
2.Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. **Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses.** Statistics in medicine. 2000;19(13):1707-28.
3.Rhodes KM, Turner RM, Higgins JP. **Empirical evidence about inconsistency among studies in a pair-wise meta-analysis.** Res Synth Methods. 2016;7(4):346-70.
4.White IR, Elbourne D. **Assessing subgroup effects with binary data: can the use of different effect measures lead to different conclusions?** BMC medical research methodology. 2005;5:15.
5.VanderWeele TJ. **On the distinction between interaction and effect modification.** Epidemiology. 2009;20(6):863-71.
6.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(1):93-8.
7.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.
8.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(4):405-11.
9.Oxman AD, Guyatt GH. **A consumer's guide to subgroup analyses.** Annals of internal medicine. 1992;116(1):78-84.
10.Simon R. **Patient subsets and variation in therapeutic efficacy.** British journal of clinical pharmacology. 1982;14(4):473-82.
11.Matsuyama Y, Morita S. **Estimation of the average causal effect among subgroups defined by post-treatment variables.** Clinical trials (London, England). 2006;3(1):1-9.
12.Hirji KF, Fagerland MW. **Outcome based subgroup analysis: a neglected concern.** Trials. 2009;10:33.
13.Cuzick J. **The assessment of subgroups in clinical trials.** Experientia Supplementum. 1982;41:224-35.
14.Cook DI, Gebski VJ, Keech AC. **Subgroup analysis in clinical trials.** The Medical journal of Australia. 2004;180(6):289-91.
15.Desai M, Pieper KS, Mahaffey K. **Challenges and solutions to pre- and post-randomization subgroup analyses.** Current cardiology reports. 2014;16(10):531.
16.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 medical research methodology. 2017;17(1):173.
17.Grady D, Cummings SR, Hulley SB. Chapter 11: Alternative trial designs and implementation issues. In: Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB, editors. Designing Clinical Research. 3 ed. Philadelphia: LIPPINCOTT WILLIAMS & WILKINS,; 2007.
18.Moye LA. Chater 21: The multiple comparison issue in health care research. In: Rao CR, Miller JP, Rao DC, editors. Handbok of statistics: epidemiology and medical statistics. 1 ed. Amsterdam: Elsevier; 2008.
19.Rosenbaum PR. **The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment.** Journal of the Royal Statistical Society. 1984;147(4):656-66.

20.Korn EL, Othus M, Chen T, Freidlin B. **Assessing treatment efficacy in the subset of responders in a randomized clinical trial.** Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2017;28(7):1640-7.
21.Van den Berghe G, Wilmer A, Hermans G, Meersseman W, Wouters PJ, Milants I, et al. **Intensive insulin therapy in the medical ICU.** The New England journal of medicine. 2006;354(5):449-61.
22.Dahabreh IJ, Trikalinos TA, Kent DM, Schmid CH. Heterogeneity of Treatment Effects. In: Gatsonis C, Morton SC, editors. Methods in Comparative Effectiveness Research. Boca Raton: CRC Press; 2017.
23.Henderson NC, Louis TA, Wang C, Varadhan R. **Bayesian analysis of heterogeneous treatment effects for patient-centered outcomes research.** Health Serv Outcomes Res Methodol. 2016;16(4):213-33.
24.Thompson SG. **Why sources of heterogeneity in meta-analysis should be investigated.** Bmj. 1994;309(6965):1351-5.
25.Fletcher J. **Subgroup analyses: how to avoid being misled.** Bmj. 2007;335(7610):96-7.
26.Dijkman B, Kooistra B, Bhandari M, Evidence-Based Surgery Working G. **How to work with a subgroup analysis.** Canadian journal of surgery Journal canadien de chirurgie. 2009;52(6):515-22.
27.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 medical research methodology. 2013;13:106.
28.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. pcori.org; 2012.
29.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.
30.Russell JA, Walley KR, Singer J, Gordon AC, Hebert PC, Cooper DJ, et al. **Vasopressin versus norepinephrine infusion in patients with septic shock.** The New England journal of medicine. 2008;358(9):877-87.
31.**Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2. ISIS-2 (Second International Study of Infarct Survival) Collaborative Group.** Lancet. 1988;2(8607):349-60.
32.**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.** Bmj. 1994;308(6921):81-106.
33.Parienti JJ, Thirion M, Megarbane B, Souweine B, Ouchikhe A, Polito A, et al. **Femoral vs jugular venous catheterization and risk of nosocomial events in adults requiring acute renal replacement therapy: a randomized controlled trial.** Jama. 2008;299(20):2413-22.
34.Study to Prospectively Evaluate Reamed Intramedullary Nails in Patients with Tibial Fractures I, Bhandari M, Guyatt G, Tornetta P, 3rd, Schemitsch EH, Swiontkowski M, et al. **Randomized trial of reamed and unreamed intramedullary nailing of tibial shaft fractures.** The Journal of bone and joint surgery American volume. 2008;90(12):2567-78.

35.Investigators S, Bhandari M, Guyatt G, Tornetta P, 3rd, Schemitsch E, Swiontkowski M, et al. **Study to prospectively evaluate reamed intramedually nails in patients with tibial fractures (S.P.R.I.N.T.): study rationale and design.** BMC musculoskeletal disorders. 2008;9:91.
36.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.
37.Trick WE, Miranda J, Evans AT, Charles-Damte M, Reilly BM, Clarke P. **Prospective cohort study of central venous catheters among internal medicine ward patients.** Am J Infect Control. 2006;34(10):636-41.
38.Corwin HL, Gettinger A, Fabian TC, May A, Pearl RG, Heard S, et al. **Efficacy and safety of epoetin alfa in critically ill patients.** The New England journal of medicine. 2007;357(10):965-76.
39.Corwin HL, Gettinger A, Pearl RG, Fink MP, Levy MM, Shapiro MJ, et al. **Efficacy of recombinant human erythropoietin in critically ill patients: a randomized controlled trial.** Jama. 2002;288(22):2827-35.
40.Ebbeling CB, Leidig MM, Feldman HA, Lovesky MM, Ludwig DS. **Effects of a low-glycemic load vs low-fat diet in obese young adults: a randomized trial.** Jama. 2007;297(19):2092-102.
41.Pittas AG, Das SK, Hajduk CL, Golden J, Saltzman E, Stark PC, et al. **A low-glycemic load diet facilitates greater weight loss in overweight adults with high insulin secretion but not in overweight adults with low insulin secretion in the CALERIE Trial.** Diabetes care. 2005;28(12):2939-41.
42.VanderWeele TJ, Knol MJ. **Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions.** Annals of internal medicine. 2011;154(10):680-3.
43.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 technology assessment (Winchester, England). 2001;5(33):1-56.
44.Rockette HE, Caplan RJ. **Strategies for subgroup analysis in clinical trials.** Recent results in cancer research Fortschritte der Krebsforschung Progres dans les recherches sur le cancer. 1988;111:49-54.
45.Tanniou J, van der Tweel I, Teerenstra S, Roes KCB. **Estimates of subgroup treatment effects in overall nonsignificant trials: To what extent should we believe in them?** Pharmaceutical statistics. 2017;16(4):280-95.
46.Pocock SJ, Assmann SE, Enos LE, Kasten LE. **Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems.** Statistics in medicine. 2002;21(19):2917-30.
47.Altman DG, Bland JM. **How to obtain the P value from a confidence interval.** Bmj. 2011;343:d2304.
48.Knol MJ, Pestman WR, Grobbee DE. **The (mis)use of overlap of confidence intervals to assess effect modification.** Eur J Epidemiol. 2011;26(4):253-4.
49.Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. **Redefine statistical significance.** Nature Human Behaviour. 2018;2(1):6-10.

50.Wilt TJ, Jones KM, Barry MJ, Andriole GL, Culkin D, Wheeler T, et al. **Follow-up of Prostatectomy versus Observation for Early Prostate Cancer.** The New England journal of medicine. 2017;377(2):132-42.
51.Wallentin L, Becker RC, Budaj A, Cannon CP, Emanuelsson H, Held C, et al. **Ticagrelor versus clopidogrel in patients with acute coronary syndromes.** The New England journal of medicine. 2009;361(11):1045-57.
52.collaborators C-, Roberts I, Shakur H, Afolabi A, Brohi K, Coats T, et al. **The importance of early treatment with tranexamic acid in bleeding trauma patients: an exploratory analysis of the CRASH-2 randomised controlled trial.** Lancet. 2011;377(9771):1096-101, 101 e1-2.
53.Mills JL. **Data torturing.** The New England journal of medicine. 1993;329(16):1196-9.
54.Counsell CE, Clarke MJ, Slattery J, Sandercock PA. **The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis?** Bmj. 1994;309(6970):1677-81.
55.Higgins JP, Thompson SG. **Controlling the risk of spurious findings from meta-regression.** Statistics in medicine. 2004;23(11):1663-82.
56.Li G, Taljaard M, Van den Heuvel ER, Levine MA, Cook DJ, Wells GA, et al. **An introduction to multiplicity issues in clinical trials: the what, why, when and how.** International journal of epidemiology. 2017;46(2):746-55.
57.Starr JR, McKnight B. **Assessing interaction in case-control studies: type I errors when using both additive and multiplicative scales.** Epidemiology. 2004;15(4):422-7.
58.Shaffer JP. **Multiple Hypothesis-Testing.** Annu Rev Psychol. 1995;46:561-84.
59.Varadhan R, Wang SJ. **Treatment effect heterogeneity for univariate subgroups in clinical trials: Shrinkage, standardization, or else.** Biometrical journal Biometrische Zeitschrift. 2016;58(1):133-53.
60.Grouin JM, Coste M, Lewis J. **Subgroup analyses in randomized clinical trials: statistical and regulatory issues.** Journal of biopharmaceutical statistics. 2005;15(5):869-82.
61.Burke JF, Sussman JB, Kent DM, Hayward RA. **Three simple rules to ensure reasonably credible subgroup analyses.** Bmj. 2015;351:h5651.
62.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.
63.Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. **Statistics in medicine--reporting of subgroup analyses in clinical trials.** The New England journal of medicine. 2007;357(21):2189-94.
64.Knol MJ, VanderWeele TJ. **Recommendations for presenting analyses of effect modification and interaction.** International journal of epidemiology. 2012;41(2):514-20.
65.Barnett HJ, Taylor DW, Eliasziw M, Fox AJ, Ferguson GG, Haynes RB, et al. **Benefit of carotid endarterectomy in patients with symptomatic moderate or severe stenosis. North American Symptomatic Carotid Endarterectomy Trial Collaborators.** The New England journal of medicine. 1998;339(20):1415-25.
66.Taylor DW, Barnett HJ, Haynes RB, Ferguson GG, Sackett DL, Thorpe KE, et al. **Low-dose and high-dose acetylsalicylic acid for patients undergoing carotid endarterectomy: a randomised controlled trial. ASA and Carotid Endarterectomy (ACE) Trial Collaborators.** Lancet. 1999;353(9171):2179-84.

67.Chaillet N, Dumont A, Abrahamowicz M, Pasquier JC, Audibert F, Monnier P, et al. **A cluster-randomized trial to reduce cesarean delivery rates in Quebec.** The New England journal of medicine. 2015;372(18):1710-21.
68.Fisher DJ, Carpenter JR, Morris TP, Freeman SC, Tierney JF. **Meta-analytical methods to identify who benefits most from treatments: daft, deluded, or deft approach?** Bmj. 2017;356:j573.
69.Royston P, Altman DG, Sauerbrei W. **Dichotomizing continuous predictors in multiple regression: a bad idea.** Statistics in medicine. 2006;25(1):127-41.
70.Altman DG, Royston P. **The cost of dichotomising continuous variables.** Bmj. 2006;332(7549):1080.
71.Altman DG, Lausen B, Sauerbrei W, Schumacher M. **Dangers of using "optimal" cutpoints in the evaluation of prognostic factors.** Journal of the National Cancer Institute. 1994;86(11):829-35.
72.Royston P, Sauerbrei W. **Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis.** Statistics in medicine. 2013;32(22):3788-803.
73.Gilthorpe MS, Clayton DG. Statistical Interactions and Gene-Environment Joint Effects. In: Tu YK, Greenwood DC, editors. Modern methods for Epidemiology. 1 ed. Dordrecht: Springer; 2012.
74.Royston P, Sauerbrei W. **Interaction of treatment with a continuous variable: simulation study of power for several methods of analysis.** Statistics in medicine. 2014;33(27):4695-708.
75.Royston P, Sauerbrei W. **A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials.** Statistics in medicine. 2004;23(16):2509-25.
76.Bonetti MZ, D.; Cole, B. F.; Gelber, R. D. **A small sample study of the STEPP approach to assessing treatment-covariate interactions in survival data.** Statistics in medicine. 2009;28(8):1255-68.
77.The CRASH Trials Co-ordinating Centre. **Protocol 05PRT/1:The CRASH-2 (Clinical Randomization of an Anti-fibrinolytic in Significant Haemorrhage) trial.** Lancet. 2005;avaialble from: https://www.thelancet.com/protocol-reviews/05PRT-1.
78.Medical Research Council Renal Cancer Collaborators. **Interferon-alpha and survival in metastatic renal carcinoma: early results of a randomised controlled trial.** Lancet. 1999;353(9146):14-7.
79.Royston P, Sauerbrei W, Ritchie A. **Is treatment with interferon-alpha effective in all patients with metastatic renal carcinoma? A new approach to the investigation of interactions.** British journal of cancer. 2004;90(4):794-9.
80.VanderWeele TJ. Explanation in causal inference. Methods for mediation and interaction. 1 ed. New York: Oxford University Press; 2015.
81.Pearce N, Greenland S. Confounding and Interaction In: Ahrens W, Pigeot I, editors. Handbook of Epidemiology. 2 ed. New York: Springer Science + Business Media; 2014.
82.Varadhan R, Wang SJ. **Standardization for subgroup analysis in randomized controlled trials.** Journal of biopharmaceutical statistics. 2014;24(1):154-67.

83.Higgins J, Sterne J, Savović J, Page M, Hróbjartsson A, Boutron I, et al. A revised tool for assessing risk of bias in randomized trials. In: Chandler J, McKenzie J, Boutron I, Welch V, editors. Cochrane Methods: Cochrane Database of Systematic Reviews 2016;(10 Suppl 1); 2016.
84.Sun XB, M.; Busse, J. W.; You, J. J.; Akl, E. A.; Mejza, F.; Bala, M. M.; Bassler, D.; Mertz, D.; Diaz-Granados, N.; Vandvik, P. O.; Malaga, G.; Srinathan, S. K.; Dahm, P.; Johnston, B. C.; Alonso-Coello, P.; Hassouneh, B.; Truong, J.; Dattani, N. D.; Walter, S. D.; Heels-Ansdell, D.; Bhatnagar, N.; Altman, D. G.; Guyatt, G. H. **The influence of study characteristics on reporting of subgroup analyses in randomised controlled trials: systematic review.** Bmj. 2011;342:d1569.
85.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.** European journal of cancer. 2015;51(18):2732-9.
86.Gabler NB, Duan N, Raneses E, Suttner L, Ciarametaro M, Cooney E, et al. **No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals.** Trials. 2016;17(1):320.
87.Alosh M, Huque MF, Bretz F, D'Agostino RB, Sr. **Tutorial on statistical considerations on subgroup analysis in confirmatory clinical trials.** Statistics in medicine. 2017;36(8):1334-60.
88.Thompson SG, Higgins JP. **How should meta-regression analyses be undertaken and interpreted?** Statistics in medicine. 2002;21(11):1559-73.
89.Davey Smith G, Egger M, Phillips AN. **Meta-analysis. Beyond the grand mean?** Bmj. 1997;315(7122):1610-4.
90.Berlin JA. **Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies.** American journal of epidemiology. 1995;142(4):383-7.
91.Borenstein M, Hedges L, Higgins JP, Rothstein H. Introduction to meta-analysis. 1 ed. Chichester: John Wiley & Sons; 2009.
92.Davey Smith G, Egger M. Going beyond the grand mean: subgroup analysis in meta-analysis of randommized trials. In: Egger M, Davey Smith G, Altman DG, editors. Systematic reviews in helath care: meta-analysis in context. 2 ed. London: BMJ; 2001.
93.Hingorani AD, Windt DA, Riley RD, Abrams K, Moons KG, Steyerberg EW, et al. **Prognosis research strategy (PROGRESS) 4: stratified medicine research.** Bmj. 2013;346:e5793.
94.Simmonds MC, Higgins JP. **Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data.** Statistics in medicine. 2007;26(15):2982-99.
95.Lambert PC, Sutton AJ, Abrams KR, Jones DR. **A comparison of summary patient-level covariates in meta-regression with individual patient data meta-analysis.** Journal of clinical epidemiology. 2002;55(1):86-94.
96.Berlin JA, Santanna J, Schmid CH, Szczech LA, Feldman HI, Anti-Lymphocyte Antibody Induction Therapy Study G. **Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: ecological bias rears its ugly head.** Statistics in medicine. 2002;21(3):371-87.
97.Fisher DJC, A. J.; Tierney, J. F.; Parmar, M. K. **A critical review of methods for the assessment of patient-level interactions in individual participant data meta-analysis of randomized trials, and guidance for practitioners.** Journal of clinical epidemiology. 2011;64(9):949-67.

98.Song F, Bachmann MO. **Cumulative subgroup analysis to reduce waste in clinical research for individualised medicine.** BMC medicine. 2016;14(1):197.
99.Hua HR, Burke DL, Crowther MJ, Ensor J, Smith CT, Riley RD. **One-stage individual participant data meta-analysis models: estimation of treatment-covariate interactions must avoid ecological bias by separating out within-trial and across-trial information.** Statistics in medicine. 2017;36(5):772-89.
100.Bachmann S, Finger C, Huss A, Egger M, Stuck AE, Clough-Gorr KM. **Inpatient rehabilitation specifically designed for geriatric patients: systematic review and meta-analysis of randomised controlled trials.** Bmj. 2010;340:c1718.
101.Ronellenfitsch U, Schwarzbach M, Hofheinz R, Kienle P, Kieser M, Slanger TE, et al. **Preoperative chemo(radio)therapy versus primary surgery for gastroesophageal adenocarcinoma: systematic review with meta-analysis combining individual patient and aggregate data.** European journal of cancer. 2013;49(15):3149-58.
102.Chemoradiotherapy for Cervical Cancer Meta-Analysis C. **Reducing uncertainties about the effects of chemoradiotherapy for cervical cancer: a systematic review and meta-analysis of individual patient data from 18 randomized trials.** Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2008;26(35):5802-12.
103.Bower P, Kontopantelis E, Sutton A, Kendrick T, Richards DA, Gilbody S, et al. **Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data.** Bmj. 2013;346:f540.
104.Higgins JPT, S. G.; Deeks, J. J.; Altman, D. G. **Measuring inconsistency in meta-analyses.** Bmj. 2003;327(7414):557-60.
105.Rothwell PM, Cook NR, Gaziano JM, Price JF, Belch JFF, Roncaglioni MC, et al. **Effects of aspirin on risks of vascular events and cancer according to bodyweight and dose: analysis of individual patient data from randomised trials.** Lancet. 2018;392(10145):387-99.
106.Rubio-Aparicio M, Sanchez-Meca J, Lopez-Lopez JA, Botella J, Marin-Martinez F. **Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled versus separate estimates of the residual between-studies variances.** The British journal of mathematical and statistical psychology. 2017;70(3):439-56.
107.Siemieniuk RA, Agoritsas T, Manja V, Devji T, Chang Y, Bala MM, et al. **Transcatheter versus surgical aortic valve replacement in patients with severe aortic stenosis at low and intermediate risk: systematic review and meta-analysis.** Bmj. 2016;354:i5130.
108.Schandelmaier S, Kaushal A, Lytvyn L, Heels-Ansdell D, Siemieniuk RA, Agoritsas T, et al. **Low intensity pulsed ultrasound for bone healing: systematic review of randomized controlled trials.** Bmj. 2017;356:j656.
109.Leppin AL, Gionfriddo MR, Kessler M, Brito JP, Mair FS, Gallacher K, et al. **Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials.** JAMA Intern Med. 2014;174(7):1095-107.
110.Briel M, Meade M, Mercat A, Brower RG, Talmor D, Walter SD, et al. **Higher vs lower positive end-expiratory pressure in patients with acute lung injury and acute respiratory distress syndrome: systematic review and meta-analysis.** Jama. 2010;303(9):865-73.
111.Sahgal A, Aoyama H, Kocher M, Neupane B, Collette S, Tago M, et al. **Phase 3 trials of stereotactic radiosurgery with or without whole-brain radiation therapy for 1 to 4 brain**

metastases: individual patient data meta-analysis. International journal of radiation oncology, biology, physics. 2015;91(4):710-7.
112.Schandelmaier S, Busse JW, Lytvyn L, Kaushal A, Agoritsas T, Mollon B, et al. **Low intensity pulsed ultrasound for fractures: updated systematic review of randomized controlled trials.** PROSPERO. 2016;CRD42016050965 Available from: http://www.crd.york.ac.uk/PROSPERO/display_record.php?ID=CRD42016050965.
113.Borenstein M, Higgins JP. **Meta-analysis and subgroups.** Prevention science : the official journal of the Society for Prevention Research. 2013;14(2):134-43.
114.Thompson SG, Sharp SJ. **Explaining heterogeneity in meta-analysis: a comparison of methods.** Statistics in medicine. 1999;18(20):2693-708.
115.Simmonds M, Stewart G, Stewart L. **A decade of individual participant data meta-analyses: A review of current practice.** Contemporary clinical trials. 2015;45(Pt A):76-83.
116.Wang XV, Cole B, Bonetti M, Gelber RD. **Meta-STEPP: subpopulation treatment effect pattern plot for individual patient data meta-analysis.** Statistics in medicine. 2016;35(21):3704-16.
117.Kasenda B, Sauerbrei W, Royston P, Mercat A, Slutsky AS, Cook D, et al. **Multivariable fractional polynomial interaction to investigate continuous effect modifiers in a meta-analysis on higher versus lower PEEP for patients with ARDS.** BMJ Open. 2016;6(9):e011148.
118.Schandelmaier S, Chang Y, Devasenapathy N, Devji T, Kwong JSW, Colunga Lozano LE, et al. **A systematic survey of suggested criteria for assessing the credibility of effect modification in randomized controlled trials or meta-analyses.** under review in Journal of Clinical Epidemiology.
119.Blackstone EH, Suri RM, Rajeswaran J, Babaliaros V, Douglas PS, Fearon WF, et al. **Propensity-matched comparisons of clinical outcomes after transapical or transfemoral transcatheter aortic valve replacement: a placement of aortic transcatheter valves (PARTNER)-I trial substudy.** Circulation. 2015;131(22):1989-2000.
120.Whiting P, Savovic J, Higgins JP, Caldwell DM, Reeves BC, Shea B, et al. **ROBIS: A new tool to assess risk of bias in systematic reviews was developed.** Journal of clinical epidemiology. 2016;69:225-34.
121.Hahn S, Williamson PR, Hutton JL, Garner P, Flynn EV. **Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies.** Statistics in medicine. 2000;19(24):3325-36.
122.Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. **GRADE guidelines: 7. Rating the quality of evidence--inconsistency.** Journal of clinical epidemiology. 2011;64(12):1294-302.
123.Groenwold RH, Donders AR, van der Heijden GJ, Hoes AW, Rovers MM. **Confounding of subgroup analyses in randomized data.** Archives of internal medicine. 2009;169(16):1532-4.
124.VanderWeele TJ, Robins JM. **Four types of effect modification: a classification based on directed acyclic graphs.** Epidemiology. 2007;18(5):561-8.
125.Poole C, Shrier I, VanderWeele TJ. **Is the Risk Difference Really a More Heterogeneous Measure?** Epidemiology. 2015;26(5):714-8.
126.Lesko CR, Henderson NC, Varadhan R. **Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research.** Journal of clinical epidemiology. 2018;100:22-31.

127.Schmid CH, Lau J, McIntosh MW, Cappelleri JC. **An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials.** Statistics in medicine. 1998;17(17):1923-42.

Chapter 5: Discussion

Achievements

Compared to previously available quality appraisal instruments for effect modification, such as the one presented in Chapter 2, ICEMAN represents a more transparent, and likely more reliable and valid credibility assessment approach. Table 1 summarizes the main achievements.

Table 1: Comparison of the situation before and after development of ICEMAN

	Situation before ICEMAN	Now
Concept	Previous checklists vague with respect to definition of effect modification and credibility, distinction between credibility and importance, and study design	ICEMAN makes these considerations explicit
Response options	Either not available or yes – no – unclear	Four well defined response options with examples for each response option
Additional credibility considerations	Unclear	Explicitly presented
Overall credibility	Either not available or yes – no – unclear	Visual analogue scale divided in four ordered categories very low, low, moderate, and high credibility
Expert input	Two previous checklists were based on an expert consensus with important limitations (e.g. unclear eligibility criteria)[1,2]	Expert consensus with explicit eligibility criteria and invitation of experts at random
User testing	Not done	User testing based on semi-structured interviews and a formal qualitative analysis
Compatibility with other quality appraisal tools	Unclear	Compatible, explained in manual
Reporting	Unclear	Explicit suggestions in manual
Consequences for decision making	Vague	Explicit suggestions in instrument and manual
Elaboration	One checklist included a detailed explanation document[1]	Comprehensive manual with instructions, rationales, and examples from the literature
Reliability	Unknown	Unknown; formal reliability study planned
Validity	Simulation studies support validity of statistical credibility criteria (e.g. tests of interaction, multiplicity, dependency on prior knowledge, and power)	Same situation with likely improved content validity through rigorous systematic survey and expert consensus and likely improved face validity through user-testing

New features that will likely increase the transparency of the credibility assessment include an explicit underlying concept, preliminary considerations to define the effect modification under consideration, pre-defined response options with detailed definitions (if possible using numerical thresholds), inclusion of supporting comments and quotations, and suggestion for reporting the overall credibility assessment. ICEMAN encourages users to report credibility as one of four categories very low, low, moderate, and high credibility and justify the rating by referring to the items that suggested reduced credibility. Previous approaches have encouraged users to use dichotomous and vague terminology such as *exploratory* versus *confirmatory*.[2]

Reliability increases with standardization, clarity, and variation of responses.[3] The rigorous development process, namely multiple rounds of expert input and, most importantly, a thorough user testing, improved clarity of items and instructions. To ensure sufficient variation, we sought to provide examples for all possible response options in the manual. Therefore, we are confident that ICEMAN is likely to provide considerably greater reliability than previous approaches (other approaches have not, however, been tested). At this time, however, we do not have reliability data available and are therefore planning to perform a formal reliability study.

Given that ICEMAN addresses the validity of subgroup effects, it would be interesting to know how accurate the resulting credibility ratings are or whether they are biased towards being over-critical or over-optimistic. Evaluation of validity, however, is unlikely to be possible. The fundamental problem is that no concurrent reference standard is available – the analysis of effect modification under consideration will often be the best data available. It might be possible, however, to test the predictive validity of an ICEMAN rating, i.e. by defining a references standard in the future. A potential study would be to assess the extent to which ICEMAN credibility ratings can predict the probability that an effect modification proves true or false in the long run as evidence accumulates. The study could theoretically be done retrospectively. We would first identify claims of effect modifications that have proven definitely true or definitely false, for instance if they could be consistently replicated or refuted in a number of subsequent RCTs.[4] The next step would be to select the initial claim and ask a rater who is blinded to the subsequent RCTs to apply ICEMAN. Once these steps have been completed for a large number of definitely true and false claims, one could assess the correlation of the initial ICEMAN ratings and whether the claims proved true or false. Such an approach, however, seems currently not feasible: although we know a number of examples in which claims of effect modification have later been refuted,[5,6] examples of claims that have stood the test of time are extremely rare.[7] It might be more realistic to perform a prospective study, e.g. by establishing a cohort of claims of effect modification rated by ICEMAN, and systematically collect attempts to replicate the claim until a sufficient number of claims have been established as highly credible. A possible time frame could be 20 years.

Alternative but less strong indicators of validity are the likely high content validity and face validity of ICEMAN, i.e. the extent to which the scale addresses all relevant aspects of credibility (content validity) and generally appears appropriate to measure credibility of effect modification (face validity).[3] We have completed a development approach that optimizes both

content and face validity. The rigorous systematic survey of the literature (chapter 3) and expert consensus study (chapter 4) make it very unlikely that we missed credibility considerations that other experts would deem relevant. In the user testing (chapter 4), many users spontaneously described the instrument as useful, which suggests high face validity.

In addition to improving transparency, reliability, and validity, ICEMAN may have an educational benefit: Most of the participants in the user-testing spontaneously mentioned that they found ICEMAN instructive, in particular the manual. This gives us hope that ICEMAN might to some extent influence the planning of new analyses of effect modification. Ideally, investigators who have used ICEMAN would consider the credibility items when they develop their next study protocol or analysis plan. In particular, they might more frequently consider hypotheses and prior evidence (e.g. guided by the PROGRESS framework[8]), consideration of fewer candidate effect modifiers (e.g. by combing effect modifiers in a risk score[9]), explicit definitions of candidate effect modifiers, more frequent use of interaction tests, refraining from dichotomizing continuous effect modifiers, and consideration of within-trial analyses of effect modification when performing a meta-analysis. To increase the likelihood that ICEMAN will improve planning, we plan to develop associated reporting guidelines (see below).

Potential risks associated with use of ICEMAN

Apart from the likely benefits of ICEMAN, formal quality appraisal instruments entail some risk of being misused. A potential risk of misusing analyses of effect modification are situations in which the RCT or meta-analysis suggests overall no effect. In such situations, disappointed investigators might be tempted to perform many analyses of effect modification with the hope to identify a subgroup of patients in whom the interventions might still look promising. ICEMAN is unlikely to be misused for those data-dredging exercises – in contrast, it will likely suggest very low credibility.

The only potential situation for misuse that we see is when investigators have an interest in showing that a potentially true effect modification does not exist. They could artificially lower the credibility rating by intentionally failing to be transparent regarding available prior evidence and using inferior methods such as dichotomizing continuous variables or choosing suboptimal models for individual participant data meta-analyses. We will monitor citations of ICEMAN to investigate whether such a misuse might be a real or just a theoretical concern.

Dissemination strategy

The success of ICEMAN depends on its successful uptake in the research community. The following considerations are part of our dissemination strategy:

- We will publish ICEMAN in a journal (Chapter 4), ideally high impact
- In parallel to the journal publication, we are planning to create a separate website from which users can download ICEMAN and the guidance document in their preferred format, including potential updates. On the website, we will encourage users to share with us their experience with ICEMAN and possibly completed instruments.

- We will suggest ICEMAN for inclusion in the main methodological guidance documents for RCTs and meta-analyses. Those include the Cochrane handbook,[10] the GRADE handbook,[11] and a new library of quality appraisal tools that is currently under development.[12]
- Optimal use of ICEMAN depends on reporting. Therefore, we plan to develop related reporting guidelines. Most likely, we will develop extensions to CONSORT[13] (for RCTs), SPIRIT[14] (for protocols of RCTs) PRISMA[15] (for meta-analyses) and PRISMA-P[16] (for protocols of meta-analyses)
- We are considering to translating ICEMAN to Chinese

Prospects for future research

ICEMAN is selective in its focus on claims that an effect modification is present and its focus on RCTs and meta-analyses of RCTs.

A future project could be to develop an instrument to assess the credibility of claims of absence of effect modification. The methodological literature addressing absence of effect modification is scarce. Neither non-significant tests of interaction nor a low amount of heterogeneity in a meta-analysis are sufficient to conclude that relevant effect modification is absent.[17-19] In our systematic review of credibility considerations (Chapter 3), we did not identify criteria for assessing whether a treatment effect is consistent across levels of a potential effect modifier. Developing such an instrument might be the subject of a future project.

Another potential project would be to extend ICEMAN to non-randomized studies. Lack of randomization, however, would introduce the major concern of confounding for the main effect. For most non-randomized studies, the risk of confounding will likely greatly reduce credibility to low just by design. Because of the impact of confounding, an extended version of ICEMAN would likely have a great overlap with existing risk of bias tools for non-randomized studies such as ROBINS-I.[20]

Before we can create such an instrument, however, we would first have to clarify a related issue: We found very little information about the extent to which the risk of bias (including confounding) of the overall effect influences the credibility of a potential effect modification. (Because of the current lack of knowledge, ICEMAN includes this potentially relevant consideration only as an optional consideration).

Apart from developing other versions of ICEMAN, the following project ideas arose during the instrument development process:

- To synthesize the many meta-epidemiological studies (around 30) that addressed aspects of the quality of analyses of effect modification. One might organize such a synthesis by using the ICEMAN structure (i.e. empirical evidence on practice of formulating hypotheses, empirical evidence on practice to referring to prior knowledge, empirical evidence of use of interaction tests, and so on)
- To clarify the consequences of considering different types of effect modifiers. For example, if an effect modifier is an intervention characteristic, a causal interpretation is required,

whereas a causal interpretation is not needed if an effect modifier is a patient characteristic. In meta-analysis, is it also common to consider effect modification by methodological characteristics (e.g. risk of bias) and then, if found to be credible, discard the subgroup with inferior methodology.

- To systematically apply ICEMAN to a large number of claimed effect modifications and assess whether its conclusions support credibility hypotheses that we identified in the literature, e.g. that qualitative effect modification is generally less credible than quantitative effect modification;[21] or true effect modification is more likely in target-specific than in unspecific cancer therapy.[22]

In summary, ICEMAN is a rigorously developed quality appraisal instrument for claims of effect modification that, compared to previous checklists, includes a number of new features that are likely to improve the quality and usefulness of credibility ratings. Future projects should investigate the reliability of the credibility ratings and, to optimize dissemination, introduce the credibility considerations to key guidance documents and reporting guidelines for RCTs and meta-analyses.

References

1.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 medical research methodology. 2017;17(1):173.
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 medical research methodology. 2011;11:14.
3.Streiner DL, Norman GR, Cairney J. Health measurement scales : a practical guide to their development and use. Oxford: Oxford University Press; 2015.
4.Song F, Bachmann MO. **Cumulative subgroup analysis to reduce waste in clinical research for individualised medicine.** BMC medicine. 2016;14(1):197.
5.Rothwell PM. **Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation.** Lancet. 2005;365(9454):176-86.
6.Guyatt G. Users' guides to the medical literature : a manual for evidence-based clinical practice. New York: McGraw-Hill Education; 2015.
7.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.
8.O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. **Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health.** Journal of clinical epidemiology. 2014;67(1):56-64.
9.Hayward RAK, D. M.; Vijan, S.; Hofer, T. P. **Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis.** BMC medical research methodology. 2006;6:18.
10.Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Higgins JP, Green SB, editors: The Cochrane Collaboration; 2011.
11.Schunemann H, Brozek J, GH G, Oxman AD. GRADE Handbook 2013 [Available from: https://gdt.gradepro.org/app/handbook/handbook.html#h.hnedbo8gqjqk.
12.Whiting P, Wolff R, Mallett S, Simera I, Savovic J. **A proposed framework for developing quality assessment tools.** Systematic reviews. 2017;6(1):204.
13.Schulz KFA, D. G.; Moher, D.; Consort Group. **CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials.** Bmj. 2010;340:c332.
14.Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gotzsche PC, Krleza-Jeric K, et al. **SPIRIT 2013 statement: defining standard protocol items for clinical trials.** Ann Intern Med. 2013;158(3):200-7.
15.Liberati AA, D. G.; Tetzlaff, J.; Mulrow, C.; Gotzsche, P. C.; Ioannidis, J. P.; Clarke, M.; Devereaux, P. J.; Kleijnen, J.; Moher, D. **The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration.** PLoS medicine. 2009;6(7):e1000100.
16.Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. **Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement.** Syst Rev. 2015;4:1.

17.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 medical research methodology. 2013;13:106.
18.Groenwold RHR, M. M.; Lubsen, J.; van der Heijden, G. J. **Subgroup effects despite homogeneous heterogeneity test results.** BMC medical research methodology. 2010;10:43.
19.Keene ON, Garrett AD. **Subgroups: time to go back to basic statistical principles?** Journal of biopharmaceutical statistics. 2014;24(1):58-71.
20.Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, et al. **ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions.** BMJ. 2016;355:i4919.
21.Weiss NS. **Subgroup-specific associations in the face of overall null results: should we rush in or fear to tread?** Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008;17(6):1297-9.
22.Yusuf S, Wittes J. **Interpreting Geographic Variations in Results of Randomized, Controlled Trials.** The New England journal of medicine. 2016;375(23):2263-71.