Instrument for the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses: manual *Version 1.0*

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

This manual provides, for each element of ICEMAN, detailed explanations, key references, examples, suggestions for use and presentation, and elaboration on conceptual considerations.

Table of contents

Appendix to: Schandelmaier S, Briel M, Varadhan R, et al. Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses. *CMAJ* 2020. doi: 10.1503/cmaj.200077. Copyright © 2020 The Author(s) or their employer(s).
To receive this resource in an accessible format, please contact us at cmajgroup@cmaj.ca.

1 Preliminary considerations

The assessment starts with a set of preliminary considerations to clarify the sources of information and help define the effect modification under consideration.

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.[1-4] Therefore, the credibility rating is likely to differ depending on the chosen effect measure.

> *Example: An RCT showed that a lifestyle modification program can prevent diabetes.[5] A subgroup analysis compared the preventive effect after dividing the patient in four groups according to their predicted risk of developing diabetes. On the relative hazard ratio scale, the effect of the lifestyle modification was consistent across risk groups (i.e., no suggestion of effect modification). On the absolute risk difference scale, however, the effect was much greater in high-risk than in low-risk patients.[6]*

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., only *one* effect modifier on each ICEMAN form). Effect modifiers may be patient characteristics (e.g., disease severity, age, or type of tumor), intervention alternatives (e.g., different doses, co-interventions, or modes of administration), or, in a meta-analysis, methodological study characteristics (e.g., risk of bias, outcome definition, type of funding). Note that 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 or at randomization? The instrument applies to potential effect modifiers assessed before or immediately after randomization, e.g., baseline variables or a stratification variable at randomization. If the effect modifier was measured after randomization, e.g., an intermediate outcome, the assessment of effect modification is complicated and potentially misleading.[7-21] Those analyses require different methods[13, 17, 22] 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 such as 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.*[23] *Length of ICU stay (i.e., the apparent effect modifier), 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.*

2 ICEMAN for randomized controlled trials

2.1 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.[24, 25]

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.[8, 11, 26-30] Hypotheses are most credible if verified in a prior, ideally in a time-stamped protocol or analysis plan.

Note that statements of pre-specification may not be reliable,[31] nor do they imply a specific hypothesis, nor do they preclude issues of multiple analyses.

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.[32]

[ ] 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.[33] 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.[34]

[ ] 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.[35]

[ ] 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.[36] 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.[37]

2.2 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[38] 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 a different direction of effect modification.
Example: A trial investigated the efficacy of neratinib in women with breast cancer who had previously received trastuzumab. An analysis of effect modification suggested that neratinib provided a greater survival benefit to women with hormone receptor-positive (HR 0.51, 95% CI 0.33–0.77) than to those with hormone receptor-negative cancer (0·93, 95% CI 0.60–1.43; interaction p-value=0·054).[39] In the discussion, the authors mention two related trials that showed similar survival benefits irrespective of hormone receptor status,[40, 41] and three other related trials that suggested greater survival in patients with hormone receptor-negative disease than in those with hormone receptor-positive disease, i.e., the opposite direction of effect modification.[42-44].

[ ] **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.[35] 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.[45]

[ ] **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.[46] 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.[47]

[ ] **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).[48] A previous RCT cited in the paper found a similar, significant effect modification (interaction p=0.02).[49]

2.3 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 if the null hypothesis (i.e., no effect modification) were true.[50-52] 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.[50])

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.[11, 12, 51, 53, 54]

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).[55, 56] 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.[57] 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 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).[58] 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.[59]

[ ] 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.[36]

[ ] 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.[60]

2.4 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,[61-63] 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.[64] 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,[65] 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,[66] testing all candidate effect modifiers in a common model, using a composite variable such as a risk score, or shrinkage estimators.[53, 67] All techniques inevitably reduce power.[8, 68, 69] 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 [70-72]). 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.[31] 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.[73] 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.[74] 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).[58] *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). The effect modifier, high versus low risk pregnancy, was one of two effect modifiers specified in the study protocol. The protocol, however, was not published a priori but provided as an appendix to the main publication.*[75]

[ ] Definitely yes: Protocol available and 3 or fewer effect modifiers tested or number considered in analysis. *Example: A trial compared endarterectomy versus medical therapy in patients with carotid stenosis. A subgroup analysis suggested that patients with stenosis of less than 70 percent had no or little benefit, while patients with severe stenosis of 70 percent or more had a durable benefit from endarterectomy.*[73] *The published study protocol specified only one effect modifier (degree of stenosis with cut point 70%).*[76]

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

Item explanation: Categorizing continuous effect modifiers is common[77] but associated with a number of problems:[78, 79] 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.[78, 80]

Ordered categories (e.g., low, medium, high blood pressure) also depends on cut point definitions and are thus subject to potential arbitrariness. Using multiple ordered subgroups, however, can also strengthen a claim if they suggest a clear trend (see "dose response effect" under optional considerations below). Note that defining groups for nominal variables can also be arbitrary (e.g., locations arbitrarily grouped into Europe versus Asia) even though they do not involve cut points.

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.[81-83] 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.[84]

An alternative to use of cut points and potentially complex modelling is to consider overlapping subgroups (e.g., using a sliding window approach).[85] 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.[84]

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).
> *Example***:** A trial tested the effect of everolimus in women who had a specific type of breast cancer. *An analysis of effect modification suggested that Patients with a low chromosomal instability (CIN) score had a greater survival benefit from the drug (HR 0.39, 95% CI 0.28 to 0.54) than patients with a higher CIN score (HR 0.62, 95% CI 0.35 to 1.08; interaction p-value=0.17). The investigators dichotomized the continuous CIN score using the 75th percentile as a cut point. They state that "a search for an optimal cut point showed that the 75th percentile yielded the maximal difference in HR between high– versus low–CIN score subgroups".*[86]
>
> [ ] **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.*[58]
>
> [ ] **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).*[60] *The investigators had pre-specified the three categories in a published protocol.*[87]
>
> [ ] **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.*[88] *A subsequent analysis suggested white cell count as an effect modifier: 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 which avoided arbitrary cut points and maximized the power of the analysis. A plot of the continuous relationship provides confidence bands of the potential effect modification and suggest a dose-response relationship.*[89]

2.6 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.[90] 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.[17, 91, 92] 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.[60]

"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.[60]

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.[93] 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). [60]

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.[94] 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.[91] 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.[95-97]

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.[69, 98] 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.[69] 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 a 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.[36]

2.7 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

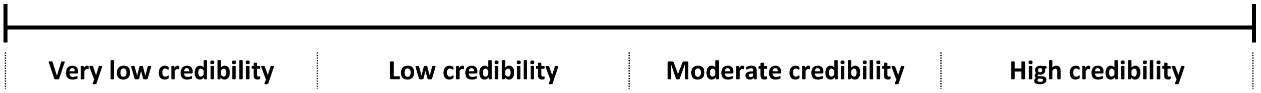
It is helpful to justify the overall rating and weighting of items using the space provided below the overall rating.

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 4 provides more suggestions for using and presenting ICEMAN in context, section 5 a 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.



Very likely no effect modification Use overall estimate for each subgroup	Likely no effect modification Use overall estimate for each subgroup but note remaining uncertainty	Likely effect modification Use separate estimates for each subgroup but note remaining uncertainty	Very likely effect modification Use separate estimates for each subgroup

Comment:

2.8 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.[60] 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: “Effects of tranexamic acid on death, vascular occlusive events, and blood transfusion in trauma patients with significant haemorrhage (CRASH-2): a randomised, placebo-controlled trial” (Lancet 2010; 376: 23–32); Secondary publication focussed on subgroup effect of interest: “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: 1096–101)

If available, protocol reference(s): available online: 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 or at 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 (not direction of effect modification) pre-specified in published protocol; explicit statement in publications that they had correctly anticipated the direction of the effect modification, although surprised by qualitative subgroup effect (i.e., benefit in some patients and harm in other patients)

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 trial or cohort study showing a similar effect modification

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). But: The p-value is very small and conclusions unlikely to be affected by multiplicity issues.

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 two analyses with consistent results: 1) based on two pre-specified cut points; 2) a continuous analysis. The authors present 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 [X] yes, probably increase

Comment: Effect modification persisted "after adjustment for interactions between the other pre-specified baseline characteristics and treatment (p<0.0001)"; The plot suggests a "dose-response" effect for the interaction

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)

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

3 ICEMAN for meta-analyses

3.1 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).[10, 77, 99-106] 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.[105-107].

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:[77, 108] 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).[77, 108]

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.[77] 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.[77, 108, 109] 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.[77, 108, 110]

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).[111] 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.[112] 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).[113] This method combines within- and between trial information and is therefore potentially affected by study-level confounding.[77]

[ ] 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.[114]

3.2 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[109]).

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[115]).

Response options and examples:

[ ] **Definitely not similar**: Effect modification reported for two or more trials and clearly different directions.
[still searching 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.[113]

[ ] **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.[114]

[ ] **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.[116]

3.3 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).[51, 63, 98, 117]

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.[118]

[ ] **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.[119]

[ ] **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.[111]

[ ] **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.[120]

3.4 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.[24, 25]

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.[8, 11, 26-30]

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 statements of pre-specification may not be reliable,[31] 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;[99] 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.[33] 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.[33] 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.[116]

[ ] 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.[121]

[ ] 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.[118]

3.5 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 if the null hypothesis (i.e., no effect modification) were true.[50-52] 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.[11, 12, 51, 53, 54]

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.[108]

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).[55, 56] 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.[57] 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 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.[112]

[ ] **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.[122]

[ ] 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.[111]

[ ] 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.[116]

3.6 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,[61-63] 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.[64] 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,[65] 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,[66] testing all candidate effect modifiers in a common model, using a composite variable such as a risk score, or shrinkage estimators.[53, 67] All techniques inevitably reduce power.[8, 68, 69] 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 [70-72]). 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.[31] 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.[120]

[ ] **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.[111] *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.
Example: An individual participant data meta-analysis assessed the effect of adding whole brain radiation therapy to stereotactic radiosurgery in patients presenting with brain metastases. The analysis suggested that age was a significant effect modifier favoring surgery alone in younger patients and no significant difference in older patients. While no protocol was available, the report includes an explicit statement that age was one of three pre-planned effect modifiers.[122]

[ ] **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.[123] *The low number of tested effect modifiers and the pre-specified definition makes multiplicity issues less likely.*[119]

3.7 Did the authors use a random effects model?

Item explanation: The credibility of claimed effect modification is higher if investigators used a *random effects model* within subgroups, i.e., allow true effects to differ among studies within subgroups: such a model allows generalization of the results beyond the studies included in the meta-analysis; this is almost always the model that we should be using.[124, 125]

The credibility is lower if investigators used **a)** a *common effect* (also called *fixed effect*, singular) model: such a model implies that, within subgroups, all studies are based on the same population and are identical to each other in all material respects; that will almost never be the case for a meta-analysis of studies identified in the literature; [124, 125] or **b)** a *fixed effects model*, which is computationally identical to a common effect model but has a different interpretation: such a model implies that the results of a subgroup will only apply to the studies included in the subgroup but cannot be generalized beyond them; this would subvert the goals of most analyses. [124, 125] Therefore, those models do not appropriately address uncertainty due to heterogeneity between studies. [124, 125]

Simulation studies have shown that failure to assume random effects increases the risk of false positive claims for both for study level and individual participant level meta-analysis.[63, 110, 117] A random effects model strengthens a test of interaction because a significant result is usually harder to achieve.[63, 99, 102, 124, 126]

If investigators state that they used a *mixed effects model* without further specification, it usually implies that they (appropriately) used a random effects model for between-study differences within subgroups and a fixed effects model for between-subgroup differences (the latter being appropriate as well[102, 124, 125]). Therefore, the appropriate answer is usually definitely yes.

The question also applies to individual-participant data meta-analysis for which an empirical study has shown that most do not apply a random effects model.[127]

Response options and examples:

[ ] Definitely no: Fixed (or common) effect or fixed effects model explicitly stated.
Example: An individual participant data meta-analysis of using aspirin for primary prevention of cardiovascular events found a significant interaction with body weight and age. The authors explicitly state that they used a fixed effects model.[116]

[ ] Probably no or unclear: Probably no random effects model or unclear.
Example: A individual participant data meta-analysis combined 13 studies 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 did not explicitly report how they modelled between-study differences. Because they used a fixed effect model for the overall analysis, it is most likely that they also used a fixed effect model within subgroups.[113]

[ ] Probably yes: Probably random (or mixed) effects model.
[still searching 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 substantially larger improvements in function than patients undergoing geriatric rehabilitation. A meta-regression analysis showed an interaction p-value of 0.01. In the methods section, the authors explicitly specified a random effects model for between study differences.[111]

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

Item explanation: Categorizing continuous effect modifiers is common[77] but associated with a number of problems:[78, 79] 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.[78, 80]

"Ordered categories (e.g., low, medium, high blood pressure) also depends on cut point definitions and are thus subject to potential arbitrariness. Using multiple ordered subgroups, however, can also strengthen a claim if they suggest a clear trend (see "dose response effect" under optional considerations below). Note that defining groups for nominal variables can also be arbitrary (e.g., locations arbitrarily grouped into Europe versus Asia) even though they do not involve cut points"

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.[81-83] 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.[84]

An alternative to use of cut points and potentially complex modelling is to consider overlapping subgroups (e.g., using a sliding window approach).[85] 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).[128, 129] 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 searching 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. [120]

[ ] **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).[111] *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 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.[129] *A previous analysis dichotomized the effect modifier and could not reveal the potential non-linear relationship.*[121] *The investigators also provided plots of the proposed effect modification including confidence limits (suggesting high uncertainty in this case).*[129]

3.9 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.[90] 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.[17, 91, 92] 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 stricter 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.[119]

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.[130]

“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, $p=0.017$).[113]

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[94] and the ROBIS tool for systematic reviews.[131] 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.[91] 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.[132] 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.[95-97]
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.[121]

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.[69, 98] 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.[69] 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.[93] 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.[116]

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.[118]

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

<u>Item explanation:</u> 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 4 provides more suggestions for using and presenting ICEMAN in context, section 5 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.

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

Comment:

3.11 Completed 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): “Reducing Uncertainties About the Effects of Chemoradiotherapy for Cervical Cancer: A Systematic Review and Meta-Analysis of Individual Patient Data From 18 Randomized Trials” (J Clin Oncol 2008 26:5802-5812)

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 categories)

Was the effect modifier measured before or at 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, which is a potential limitation. Nevertheless, the suggested effect modification is likely mostly 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: Effect modification analysis within individual trials not shown

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 is a rather large number. This reduces the risk of trial-level confounding

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 performed; no published protocol; potential multiplicity issues are a concern

7: Did the authors use a random effects model?

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

Comment: Not explicitly stated; authors used a fixed effect 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 [X] yes, probably increase

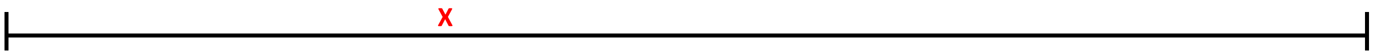
Comment: Possible “dose-response effect”; effect modification consistent across different outcomes

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)



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: Consistency across studies unclear. Prior knowledge unclear. P-value not very small and possibly inflated by multiple analyses and use of fixed effect model.

4 Practical considerations

4.1 Assessment in duplicate

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

4.2 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*).

4.3 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[94] or the ROBIS tool for meta-analyses.[131] 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[133]:

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.

5 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.[38]

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.[7, 50, 134, 135] 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.[7, 134]

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.[7]

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 (e.g., whether the outcome is patient-important;[136] the intervention results in a net benefit when considering multiple outcomes;[137] or the analysis is appropriate for the research question of interest[138]). 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.[139, 140] 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, 141] 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,[140,149] 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).[141] 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.[67]

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.

In special cases, analyses of effect modification performed in multi-center RCTs can be conceptually similar to analyses of effect modification performed in meta-analyses. Similar to study differences that can confound apparent patient-level differences in a meta-analysis, differences between centers can confound apparent patient-level differences in multi-center RCTs. In addition, similar to a random effects model applied in a meta-analysis to allow for variation of effects across studies, investigators of a multi-center trial might use a random effect model to allow for variation across centers. If centers differ substantially, and confounding is a concern, it might therefore be helpful to consider an adapted meta-analysis version of ICEMAN to claims of effect modification made in multi-center trials in which each center is treated as a trial. In general, however, the RCT version should be sufficient because the risk of center-confounding in an RCT is likely to be much lower than the risk of study-confounding in a meta-analysis; centers in the context of an RCT are usually much more similar than studies included in a meta-analysis.

On the choice of different types of random effects models: Simulation studies have shown that use of a fixed effect model is associated with a higher risk finding spurious effect modification.[63, 110, 117] Question 7 in the meta-analysis version of ICEMAN addresses this concern. Another potentially relevant consideration currently not included in ICEMAN is whether the specific sub-type of random effects model is more or less appropriate. Some types of random effects approaches can be flawed depending on the particular application. Recent publications provide preliminary guidance about the choice of model, [142] in particular with respect to issues related to meta-analyses of a small number of studies,[125, 143-148] but also state that more research is needed before clear recommendations can be made.

6 References

1. 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.
2. Venekamp RPR, M. M.; Hoes, A. W.; Knol, M. J. Subgroup analysis in randomized controlled trials appeared to be dependent on whether relative or absolute effect measures were used. Journal of clinical epidemiology. 2014;67(4):410-5.
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. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. The New England journal of medicine. 2002;346(6):393-403.
6. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. Bmj. 2018;363:k4245.
7. VanderWeele TJ. On the distinction between interaction and effect modification. Epidemiology. 2009;20(6):863-71.
8. 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.
9. 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.
10. 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.
11. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Annals of internal medicine. 1992;116(1):78-84.
12. Simon R. Patient subsets and variation in therapeutic efficacy. British journal of clinical pharmacology. 1982;14(4):473-82.
13. 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.
14. Hirji KF, Fagerland MW. Outcome based subgroup analysis: a neglected concern. Trials. 2009;10:33.
15. Cuzick J. The assessment of subgroups in clinical trials. Experientia Supplementum. 1982;41:224-35.
16. Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. The Medical journal of Australia. 2004;180(6):289-91.
17. Desai M, Pieper KS, Mahaffey K. Challenges and solutions to pre- and post-randomization subgroup analyses. Current cardiology reports. 2014;16(10):531.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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.

25. 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.

26. Thompson SG. Why sources of heterogeneity in meta-analysis should be investigated. Bmj. 1994;309(6965):1351-5.

27. Fletcher J. Subgroup analyses: how to avoid being misled. Bmj. 2007;335(7610):96-7.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

39. Chan A, Delaloge S, Holmes FA, Moy B, Iwata H, Harvey VJ, et al. Neratinib after trastuzumab-based adjuvant therapy in patients with HER2-positive breast cancer (ExteNET): a multicentre, randomised, double-blind, placebo-controlled, phase 3 trial. The Lancet Oncology. 2016;17(3):367-77.

40. Perez EA, Romond EH, Suman VJ, Jeong JH, Sledge G, Geyer CE, Jr., et al. Trastuzumab plus adjuvant chemotherapy for human epidermal growth factor receptor 2-positive breast cancer: planned joint analysis of overall survival from NSABP B-31 and NCCTG N9831. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2014;32(33):3744-52.

41. Untch M, Gelber RD, Jackisch C, Procter M, Baselga J, Bell R, et al. Estimating the magnitude of trastuzumab effects within patient subgroups in the HERA trial. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO. 2008;19(6):1090-6.

42. de Azambuja E, Holmes AP, Piccart-Gebhart M, Holmes E, Di Cosimo S, Swaby RF, et al. Lapatinib with trastuzumab for HER2-positive early breast cancer (NeoALTTO): survival outcomes of a randomised, open-label, multicentre, phase 3 trial and their association with pathological complete response. The Lancet Oncology. 2014;15(10):1137-46.

43. De Laurentiis MA, G.; Massarelli, E.; Ruggiero, A.; Carlomagno, C.; Ciardiello, F.; Tortora, G.; D'Agostino, D.; Caputo, F.; Cancello, G.; Montagna, E.; Malorni, L.; Zinno, L.; Lauria, R.; Bianco, A. R.; De Placido, S. A meta-analysis on the interaction between HER-2 expression and response to endocrine treatment in advanced breast cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2005;11(13):4741-8.

44. Dowsett M, Harper-Wynne C, Boeddinghaus I, Salter J, Hills M, Dixon M, et al. HER-2 amplification impedes the antiproliferative effects of hormone therapy in estrogen receptor-positive primary breast cancer. Cancer research. 2001;61(23):8452-8.

45. 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.

46. 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.

47. 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.

48. 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.

49. 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.

50. VanderWeele TJ, Knol MJ. Interpretation of subgroup analyses in randomized trials: heterogeneity versus secondary interventions. Annals of internal medicine. 2011;154(10):680-3.

51. 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.

52. 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.
53. 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.
54. 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.
55. Altman DG, Bland JM. How to obtain the P value from a confidence interval. Bmj. 2011;343:d2304.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. Mills JL. Data torturing. The New England journal of medicine. 1993;329(16):1196-9.
62. 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.
63. Higgins JP, Thompson SG. Controlling the risk of spurious findings from meta-regression. Statistics in medicine. 2004;23(11):1663-82.
64. 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.
65. 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.
66. Shaffer JP. Multiple Hypothesis-Testing. Annu Rev Psychol. 1995;46:561-84.
67. 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.
68. 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.
69. Burke JF, Sussman JB, Kent DM, Hayward RA. Three simple rules to ensure reasonably credible subgroup analyses. Bmj. 2015;351:h5651.
70. 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.
71. 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.
72. Knol MJ, VanderWeele TJ. Recommendations for presenting analyses of effect modification and interaction. International journal of epidemiology. 2012;41(2):514-20.
73. 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.

74. 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.

75. 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.

76. North American Symptomatic Carotid Endarterectomy Trial. Methods, patient characteristics, and progress. Stroke; a journal of cerebral circulation. 1991;22(6):711-20.

77. 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.

78. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in medicine. 2006;25(1):127-41.

79. Altman DG, Royston P. The cost of dichotomising continuous variables. Bmj. 2006;332(7549):1080.

80. 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.

81. 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.

82. 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.

83. 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.

84. 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.

85. 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.

86. Hortobagyi GN, Chen D, Piccart M, Rugo HS, Burris HA, 3rd, Pritchard KI, et al. Correlative Analysis of Genetic Alterations and Everolimus Benefit in Hormone Receptor-Positive, Human Epidermal Growth Factor Receptor 2-Negative Advanced Breast Cancer: Results From BOLERO-2. Journal of clinical oncology : official journal of the American Society of Clinical Oncology. 2016;34(5):419-26.

87. 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.

88. 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.

89. 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.

90. Schandelmaier S, Chang Y, Devasenapathy N, Devji T, Kwong JSW, Colunga Lozano LE, et al. A systematic survey identified 36 criteria for assessing effect modification claims in randomized trials or meta-analyses. Journal of clinical epidemiology. 2019;113:159-67.

91. VanderWeele TJ. Explanation in causal inference. Methods for mediation and interaction. 1 ed. New York: Oxford University Press; 2015.

92. 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.

93. Varadhan R, Wang SJ. Standardization for subgroup analysis in randomized controlled trials. Journal of biopharmaceutical statistics. 2014;24(1):154-67.

94. 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.
95. 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.
96. 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.
97. 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.
98. 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.
99. Thompson SG, Higgins JP. How should meta-regression analyses be undertaken and interpreted? Statistics in medicine. 2002;21(11):1559-73.
100. Davey Smith G, Egger M, Phillips AN. Meta-analysis. Beyond the grand mean? Bmj. 1997;315(7122):1610-4.
101. Berlin JA. Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies. American journal of epidemiology. 1995;142(4):383-7.
102. Borenstein M, Hedges L, Higgins JP, Rothstein H. Introduction to meta-analysis. 1 ed. Chichester: John Wiley & Sons; 2009.
103. 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.
104. 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.
105. 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.
106. 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.
107. 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.
108. 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.
109. Song F, Bachmann MO. Cumulative subgroup analysis to reduce waste in clinical research for individualised medicine. BMC medicine. 2016;14(1):197.
110. 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.
111. 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.

112. 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.

113. 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.

114. 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.

115. Higgins JPT, S. G.; Deeks, J. J.; Altman, D. G. Measuring inconsistency in meta-analyses. Bmj. 2003;327(7414):557-60.

116. 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.

117. 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.

118. 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.

119. 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.

120. 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.

121. 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.

122. 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.

123. 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.

124. Borenstein M, Higgins JP. Meta-analysis and subgroups. Prevention science : the official journal of the Society for Prevention Research. 2013;14(2):134-43.

125. Borenstein M. Common mistakes in meta-analysis and how to avoid them: Biostat Inc.; 2019. 388 p.

126. Thompson SG, Sharp SJ. Explaining heterogeneity in meta-analysis: a comparison of methods. Statistics in medicine. 1999;18(20):2693-708.

127. 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.

128. 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.

129. 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.

130. 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.

131. 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.

132. 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.

133. 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.

134. 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.

135. VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology. 2007;18(5):561-8.

136. Methodology Committee of the Patient-Centered Outcomes Research I. Methodological standards and patient-centeredness in comparative effectiveness research: the PCORI perspective. Jama. 2012;307(15):1636-40.

137. Alper BS, Oettgen P, Kunnamo I, Iorio A, Ansari MT, Murad MH, et al. Defining certainty of net benefit: a GRADE concept paper. BMJ Open. 2019;9(6):e027445.

138. European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. 2017.

139. Poole C, Shrier I, VanderWeele TJ. Is the Risk Difference Really a More Heterogeneous Measure? Epidemiology. 2015;26(5):714-8.

140. 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.

141. 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.

142. Veroniki AA, Jackson D, Bender R, Kuss O, Langan D, Higgins JPT, et al. Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. Res Synth Methods. 2019;10(1):23-43.

143. Michael H, Thornton S, Xie M, Tian L. Exact inference on the random-effects model for meta-analyses with few studies. Biometrics. 2018.

144. Roever C, Knapp G, Friede T. Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. BMC medical research methodology. 2015;15:99.

145. Bender R, Friede T, Koch A, Kuss O, Schlattmann P, Schwarzer G, et al. Methods for evidence synthesis in the case of very few studies. Res Synth Methods. 2018;9(3):382-92.

146. Friede T, Roever C, Wandel S, Neuenschwander B. Meta-analysis of two studies in the presence of heterogeneity with applications in rare diseases. Biometrical journal Biometrische Zeitschrift. 2017;59(4):658-71.

147. Friede T, Roever C, Wandel S, Neuenschwander B. Meta-analysis of few small studies in orphan diseases. Res Synth Methods. 2017;8(1):79-91.

148. Seide SE, Roever C, Friede T. Likelihood-based random-effects meta-analysis with few studies: empirical and simulation studies. BMC medical research methodology. 2019;19(1):16.

149. Lesko CR, Henderson NC, Varadhan R. Considerations when assessing heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol. 2018; 100:22-31.