GRADEing the quality of evidence on safety: an adapted GRADE approach for preparing lists of potentially inappropriate medication for older adults

Tim Mathes (tim.mathes@med.uni-goettingen.de)
Witten/Herdecke University
Nina-Kristin Mann
Witten/Herdecke University
Petra Thürmann
Witten/Herdecke University
Andreas Sönnichsen
Medical University of Vienna
Dawid Pieper
Witten/Herdecke University

Research Article

Keywords: Evidence synthesizes, systematic reviews, PIM-List, GRADE, Level of Evidence, safety, harms

Posted Date: December 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1169059/v1

Abstract

Background

Systematic reviews that synthesize safety outcomes pose challenges (e.g. rare events), which poses questions for grading the strength of the body of evidence.

In this contribution, we suggest adaption of the GRADE system for grading the quality of evidence on safety outcomes for developing a potentially inappropriate medication list (PRISCUS).

Methods

We systematically assessed each of the five GRADE domains for rating-down (study limitations, imprecision, inconsistency, indirectness, publication bias) and the criteria for rating-up, considering if special considerations or revisions of the original approach were indicated. The result was gathered in a written document and discussed in a group-meeting. Subsequently, we performed a proof-of-concept application using a convenience sample of systematic reviews.

Results

We adapted aspects of the criteria study limitations, imprecision, publication bias and rating-up for large effect. In addition, we suggest a new criterion to account for data from subgroup-analyses. The proof-of-concept application did not reveal a need for further revision and thus we used the approach for the systematic reviews that were prepared for the PRISCUS-list.

We assessed 51 outcomes for 19 clinical questions. Each of the proposed adaptions was applied. There were neither an excessive number of low and very low ratings, nor an excessive number of high ratings, but the different methodological quality of the safety outcomes appeared to be well reflected.

Conclusion

The adaptions appear to have the potential to overcome some of the challenges when grading the methodological quality of harms and thus may be helpful for producers of evidence syntheses considering safety.

Background

Clinical practice recommendations (e.g. in guidelines) that are based on the best available evidence can improve quality of care.[1, 2] Lists of potentially inappropriate medication (PIM) name drugs which may have a negative risk-benefit-ratio in older patients, especially when safer alternatives are available.[3, 4] Internationally, a variety of PIM lists have been prepared or adapted to the local drug market in different countries. One reason for establishing PIM lists on expert ratings (e.g. Delphi surveys) is the frequent exclusion or low inclusion rate of older patients, particularly frail elderly, in pivotal clinical trials, which

represent the basis for drug approval as well a large share of the evidence in treatment guidelines. However, an often-discussed limitation of existing PIM lists is that many of them are not based on systematic reviews of the evidence, but only on expert-opinion, unsystematic literature reviews and previously published PIM lists. This might be one reason why the overlap between PIM-lists is often low.[3] To overcome this limitation, for the update of the German PIM list, namely the PRISCUS list, the participating experts were provided data from systematic reviews conducted specifically to inform the recommendations.

To grade the quality/levels of evidence for practice recommendations grading systems have been developed.[5, 6] The Grading of Recommendations Assessment, Development and Evaluation (short GRADE) system is one of the most established tools for rating the quality of evidence underlying recommendation in clinical practice guidelines.[6] GRADE rates the quality of evidence for a specific outcome across the included studies on a PICO (participants, intervention, comparison, outcome)-question. For this purpose, explicit criteria are used. These include the study design, study limitations, imprecision, inconsistency, indirectness, dose-response association and magnitude of effect. Based on these criteria the quality of evidence is classified into four levels (high, moderate, low or very low). Noticeably, in contrast to most other approaches which classify the strength of evidence on study level (e.g. RCTs are high and case reports are low level of evidence), GRADE rates the evidence for each outcome across the included studies.

Usually, the evidence is weaker for safety than for effectiveness outcomes because of inconsistent measurement, imprecision (studies are not powered for safety outcomes, rare events) and poor reporting of harms.[7, 8] Other challenges include that safety outcomes are often rare, unpredictable or require very long follow-up times to be detected. Furthermore, harms are often sub-group specific, but the relevant groups, i.e. frail elderly participants, are underrepresented in RCTs.[9, 10] For these reasons safety may not have been sufficiently assessed in randomized controlled trials (RCTs). Consequently, only including RCTs in systematic reviews considering safety might not be sufficient. To generate sufficient evidence on safety it is advisable to include non-randomized studies (NRS) and to consider sub-groups analyses .[11]

One might argue that the quality of evidence on safety is just the way it is and no specific rating criteria for quality of evidence are necessary. However, if rating proceeds in the standard way some problems may come up. First, it bears the risk that the certainty of evidence between benefit and harms is unbalanced per se because benefits tend to get higher methodological quality ratings (e.g. because only NRS are available for a very rare harm). Second, there may be the methodological problem that a difference in the quality of evidence on different safety outcomes could not be differentiated because of floor effects, i.e. all studies are classified as low or very low quality of evidence. This is because in the GRADE system evidence from NRS always starts at low quality and thus only one additional criterion for rating down (e.g. imprecision) would result in a very low quality of evidence rating.[6]

Some older evidence level classifications schemes, e.g. the oxford level of evidence use different criteria for harms and benefits, but there is no specific GRADE guidance for assessing the quality of evidence for

safety outcomes.[5]

In this contribution, we adapted the GRADE system for grading the quality of evidence on safety outcomes and report our first experience of applying this adaption for developing a PIM list.

Methods

This project was part of the update of a PIM list for elderly patients in Germany, namely the PRISCUS list.[12] GRADE is the most established tool to assess the quality of evidence. We decided to adapt it instead of developing our own criteria for the following reasons. First, it can be assumed that in general the criteria relevant for assessing the quality of evidence on harms are (almost) the same as for benefits. Second, it facilitates the integration of evidence on benefits and harms within one evidence synthesis product (e.g. clinical practice guideline). Safety outcomes are almost always (expressed as) binary variables. Therefore, we only consider binary variables in this work.

The research team (two experienced methodologists, one senior clinical pharmacologist, one senior general practitioner, and one pharmacist) systematically assessed each of the five GRADE domains for rating down (study limitations, imprecision, inconsistency, indirectness, publication bias) and the criteria for rating up, considering whether special considerations or revisions of the original GRADE approach were indicated. A revision of an original GRADE criterion was only made if it could be methodologically justified. We judged an adaption as indicated, if it could be expected that the original GRADE criteria are affected by one of the challenges quoted in the introduction (e.g. higher imprecision, inclusion of NRS). We judged an adaption as acceptable if it could be justified based on methodological/epidemiological reasoning or could be supported by the methodological literature. The result was gathered in a written document (TM, DP) and discussed and discussed in a group meeting with the whole project team. To facilitate the discussion, we illustrated the different challenges using example cases. If necessary, we refined the criteria until a consensus was reached. Subsequently, we performed a proof-of-concept application using a convenience sample of systematic reviews focusing on safety for which we assumed that all adaptions would come into effect. These systematic reviews were not part of PRISCUS. In this phase, we checked our approach for any problems with a focus on inconsistencies and tendencies of overestimating the strength of evidence. In the proof-of-concept application no reasons for revision of the adapted criteria were recognized.

After this development phase, we used our approach for evidence syntheses prepared as basis for expert-rated recommendations on the PRISCUS-list. In the pilot study, we assessed 51 outcomes for 19 clinical questions from 13 systematic reviews.

Results

Adaptions

Table 1 shows the original GRADE criteria and the modified criteria. We explain and justify the modifications in the following text. All domains/criteria not quoted in Table 1 were applied in the usual way.

Table 1
overview of adaptions

GRADE criteria	Original/usually	Challenge	Adaptions
Study type/methodological quality	NRS start as “low quality” of evidence	Data on harms from RCTs is often insufficient and thus it is advisable to consider NRS	NRS start as “high quality” of evidence
Imprecision (binary outcomes)	Usually, 95% of CIs of relative effects are used 95%CI overlaps decision threshold (e.g. null effect) → rating down one level 95%CI includes appreciable harm and benefit → rating down two levels	Harms are often rare events and rare in the included studies despite large sample sizes. In the case of rare events 95% CIs of relative effects can be misleading.	Imprecision is assessed based on absolute effects
Publication bias/missing results in the synthesis	Rating down for publication bias	For harms selective dissemination would result in underestimation of harms	Rating up for publication bias
Large magnitude of effect	Rating up if RR >2 (<0.5)	Harms are usually less affected by confounding by indication	Rating up if RR >1.67 (<0.60)
Originally not used	Subgroup effects	Harms are often subgroup-specific but analysis within subgroups is underpowered	Rating up if there is a statistically significant subgroup effect from a well-designed subgroup analysis

Study type/methodological quality

Usually NRS start as low quality of evidence because of the risk of confounding bias.[13] The approach can be also interpreted as NRS are rated down two levels directly at the beginning because of confounding. Consequently, all evidence on harms from NRS would start at “low quality” of evidence.

Recently a new approach was suggested by the GRADE working group when using ROBINS-I for assessing NRS.[14] It suggests that NRS start high but are usually rated down by one or two levels because usually there is a risk of confounding and selection bias in NRS. Likewise, we modified the

original criteria. We suggest letting NRS also start high and then to assess if the studies suffer from confounding or participant selection bias. Furthermore, in the case that other bias in addition to confounding and selection bias are present it is possible to rate down three levels for study limitations (risk of bias).[15] Clearly this adaption is not necessary when using ROBINS-I but it can be applied when another tool or previous systematic reviews that apply other tools are used for assessing NRS. The approach appears reasonable for two reasons. First, usually confounding by indication is a minor issue for assessing harms compared to assessing effectiveness. In particular, rare harms and harms that are not obviously related to the intervention would mean that the effect is biased towards the null.[16] Second, most tools for assessing NRS consider cofounding and selection bias.[17, 18] Therefore, for most tools it is not necessary to start the assessment at low quality of evidence because this will happen in the “study limitation” domain and if the tool does not cover one of the domains, it could be assessed separately. An advantage that comes along with this approach is that double counting of confounding and selection bias is avoided, which exists if the NRS start low and the applied tool covers confounding or selection bias.

For example, in our review on safety of tramadol in the elderly, all evidence on falls risk was observational and consequently would have started at low quality of evidence, if assessed in the original way. The effect was rated-down one level for confounding bias and one level for imprecision, which would have resulted in a very low quality of evidence rating. Noticeable, in this case even only one criterion for rating-down would have led to a very low rating. In contrast, if assessed in the adapted way the evidence starts high and rating-down two levels resulted in a low quality of evidence rating.

Imprecision

Usually, precision of an effect is assessed based on the 95% CI.[19] However, in the case of rare events this might be misleading. The GRADE working group suggests using absolute effects for very low event rates, whereby “very low” is not defined. Most harms, in particular severe harms are usually rare. In addition, no fixed threshold can be determined when a 95%CI of a relative effect might be misleading. Therefore, our preliminary suggestion is to use the absolute effects for assessing harms. Important to note, the 95%CI for the absolute effect is usually not reported and must be calculated. Furthermore, in the case of harms it is especially important to use prevalence or incidence data which is applicable to the target population (e.g. elderly people) because absolute effects often vary between groups (e.g. falls in elderly people). Therefore, often external data (not from the included studies) for calculating the control group risk is preferable.

For example, in a systematic review of artemethere lumefantrine versus Amodiaquine plus sulfadoxinee pyrimethamine for treating uncomplicated malaria the authors found 34 severe adverse events (SAEs) and calculated a risk ratio (RR) of 1.08 (95%-KI 0.56 to 2.08). The SAEs were observed in in >2.700 participants resulting in an absolute risk difference of 1% with (95%CI 0.6% less to 1.4% more).[19] The example illustrates that although the 95%CIs of the RR suggest rating down two levels, the lower 95CI (possible avoidance of SAEs) of the absolute effect suggest rating down at most one level.

Publication bias

Usually, the quality of evidence is rated down one level if publication bias is detected.[20] The criteria for considering rating down because of publication bias are small industry sponsored studies and an asymmetric funnel-plot. A meta-epidemiological study found many clinical questions which were suspicious for selective dissemination of safety outcomes.[21] In contrast to benefit outcomes where publication bias would lead to an overestimation of the benefit, missing safety results would result in an underestimation of harms.[22] Therefore, we suggest that authors may consider rating-up the quality of evidence when there is strong suspicion for publication bias.

For example, in one of our reviews performed to inform experts of the panel for the PRISCUS 2.0 list on oral anticoagulants in elderly people all studies were industry sponsored. The expected harms were not reported for all studies (e.g. all bleeding events), or harms were grouped uncommonly (e.g. clinical relevant non-major bleeding), and the funnel-plot was slightly asymmetric. Therefore, we rated-up the quality of evidence for bleeding one level, from low to moderate.

Large magnitude of effect

Usually the evidence can be rated-up by one level if the magnitude of the effect is large, whereby large is defined as an RR of >2 (<0.5). This threshold is based on the assumption that confounding alone is unlikely to cause such an effect.[23] This threshold was determined based on a modelling study which was informed by older, unadjusted observational studies.[24] Newer studies suggest that if the analyses are adjusted for relevant confounders, it is unlikely that unmeasured and residual confounding lead to large or very large effects.[25, 26] Additionally, as mentioned above, harm outcomes are regularly less affected by bias by indication than effectiveness outcomes.[16] For these reasons we suggest to consider rating-up harm outcomes one level if the RR is > 1.67 (<0.60) for evidence based on unbiased and sufficiently precise NRS.[25] Moreover, we suggest to consider rating-up two levels if the RR is larger than 10 (<0.1) in NRS that are affected by confounding because previous studies suggest that such a large effect is unlikely to be caused by confounding alone.[27]

For example, in our review on protone pump inhibitors compared to no protone pump inhibitors, we found an OR of 1.97 (95%CI 1.44 to 2.70) for dementia. The risk of bias for this estimate was low. Therefore, we rated-up the quality of evidence one level, from low to moderate, which would not have been done if applied in the original way.

Subgroup-effects

Risk of experiencing harms often varies between subgroups.[28] The GRADE guidance only considers subgroup-analyses to explore inconsistency between studies.[29] However, a statistically significant test of heterogeneity not only suggests that the effects are different between different groups and consequently should be considered separately but also comprises information on the certainty of effects. If there is a statistically significant sub-group effect one could be more confident that there is an effect in one subgroup. Similarly, to a dose-response relationship, this is in particular true if the subgroup effect

increases/decreases with level of the subgroup variable (e.g. the risk of experience harms increases with age). We suggest that if a well performed subgroup analysis suggests a larger effect in one subgroup, the quality of evidence in this subgroup might be rated up one level, in particular if the subgroup effect is level dependent.[30]

For example, in our review on oral anticoagulants in elderly persons we extracted data from well performed (e.g. pre-specified, based on a test for interaction) within study subgroup-analyses for major bleeding. Most of these suggested that the risk of experiencing major bleeding increases with increasing age and that the risk of major bleeding is in particularly high in the very elderly. For that reason, we rated-up the quality of evidence one level (from low to moderate) for bleeding risk in the very elderly.

Results of pilot testing

We found RCTs (only elderly or subgroups analyses of elderly) for only 9 of the 19 clinical questions. Each of the proposed adaptions was applied Nevertheless, the ratings were well balanced. We rated 14 outcomes as high quality, 7 as moderate quality, 17 as low quality and 13 as very low quality. As expected, most “high methodological quality” ratings were made for clinical questions for which RCTs were available.

Discussion And Conclusion

Systematic reviews that synthesize safety outcomes pose specific challenges (e.g. including NRS, rare events), which come along with challenges for grading the strength of the body of evidence. In this work we propose some adaptions to or specify the application of the GRADE criteria when assessing safety.

The initial results of the application for preparing a PIM list suggests that the ratings were quite well balanced. There were neither floor-effects (excessive number of low and very low ratings) nor ceiling effects (excessive number of high ratings), but the different methodological quality of the safety outcomes seems to be well reflected. The adaptions appear to have the potential to overcome some of the challenges when grading the methodological quality of harms and thus may be helpful for producers of evidence syntheses considering safety (e.g. literature for creating PIM lists, systematic reviews on drug safety after approval). Although the adaptions were developed for evidence syntheses focusing on drug safety, we think that the adaptions might also be useful for evidence syntheses in general (i.e. all evidence syntheses considering benefits and harms) because basically all systematic reviews considering harms face the same challenges when grading the quality of evidence.

The adaptions were not developed in a GRADE working group because the timeline of the project required that the evidence syntheses start immediately. Therefore, the suggested adaptions should only be regarded as a first step for stimulating further discussion and development of guidelines specifically for grading the quality of evidence on safety. Future research is desirable for developing refined GRADE guidance for evidence syntheses on harms.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests

Funding

This work was part of the project “Updating the PRISCUS list of potentially inappropriate medications for the elderly” funded by the German Federal Ministry of Education and Research (Grant Number: 01KX1812). The funder was not involved in the study.

Authors' contributions

TM: development of the initial concept, proof of concept, application of approach, writing of manuscript

NKM: input and discussion to refine approach, application of approach, revision of manuscript

PA: input and discussion to refine approach, revision of manuscript

AS: input and discussion to refine approach, revision of manuscript

DP: development of the initial concept, writing of manuscript

Acknowledgements

None.

References

1. Lugtenberg M, Burgers JS, Westert GP: **Effects of evidence-based clinical practice guidelines on quality of care: a systematic review**. *Quality and Safety in Health Care* 2009, **18**(5):385.

2. Jano E, Aparasu RR: **Healthcare Outcomes Associated with Beers' Criteria: A Systematic Review**. *Annals of Pharmacotherapy* 2007, **41**(3):438-448.

3. Motter FR, Fritzen JS, Hilmer SN, Paniz É V, Paniz VMV: **Potentially inappropriate medication in the elderly: a systematic review of validated explicit criteria**. *Eur J Clin Pharmacol* 2018, **74**(6):679-700.

4. Renom-Guiteras A, Meyer G, Thürmann PA: **The EU(7)-PIM list: a list of potentially inappropriate medications for older people consented by experts from seven European countries**. *Eur J Clin Pharmacol* 2015, **71**(7):861-875.

5. **OCEBM Levels of Evidence Working Group*. "The Oxford Levels of Evidence 2". Oxford Centre for Evidence-Based Medicine. https://www.cebm.net/index.aspx?o=5653**. In.

6. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, Norris S, Falck-Ytter Y, Glasziou P, deBeer H *et al*: **GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables**. *Journal of Clinical Epidemiology* 2011, **64**(4):383-394.

7. Hodkinson A, Kirkham JJ, Tudur-Smith C, Gamble C: **Reporting of harms data in RCTs: a systematic review of empirical assessments against the CONSORT harms extension**. *BMJ Open* 2013, **3**(9):e003436.

8. Jia P, Lin L, Kwong JSW, Xu C: **Many meta-analyses of rare events in the Cochrane Database of Systematic Reviews were underpowered**. *Journal of Clinical Epidemiology* 2021, **131**:113-122.

9. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J: **A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results**. *Trials* 2015, **16**(1):495.

10. Luo J, Eldredge C, Cho CC, Cisler RA: **Population Analysis of Adverse Events in Different Age Groups Using Big Clinical Trials Data**. *JMIR Med Inform* 2016, **4**(4):e30.

11. **Peryer G, Golder S, Junqueira D, Vohra S, Loke YK. Chapter 19: Adverse effects. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from** In.

12. Holt S, Schmiedl S, Thürmann PA: **Potentially inappropriate medications in the elderly: the PRISCUS list**. *Dtsch Arztebl Int* 2010, **107**(31-32):543-551.

13. Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, Vist GE, Falck-Ytter Y, Meerpohl J, Norris S *et al*: **GRADE guidelines: 3. Rating the quality of evidence**. *J Clin Epidemiol* 2011, **64**(4):401-406.

14. Schünemann HJ, Cuello C, Akl EA, Mustafa RA, Meerpohl JJ, Thayer K, Morgan RL, Gartlehner G, Kunz R, Katikireddi SV *et al*: **GRADE guidelines: 18. How ROBINS-I and other tools to assess risk of bias in**

nonrandomized studies should be used to rate the certainty of a body of evidence. *J Clin Epidemiol* 2019, **111**:105-114.

15. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y *et al*: **GRADE guidelines: 4. Rating the quality of evidence–study limitations (risk of bias)**. *J Clin Epidemiol* 2011, **64**(4):407-415.

16. Jüni P, Loke Y, Pigott T, Ramsay C, Regidor D, Rothstein H, Sandhu L, Santaguida P, Schünemann H, Shea B: **Risk of bias in non-randomized studies of interventions (ROBINS-I): detailed guidance**. *Br Med J* 2016.

17. Quigley JM, Thompson JC, Halfpenny NJ, Scott DA: **Critical appraisal of nonrandomized studies-A review of recommended and commonly used tools**. *J Eval Clin Pract* 2019, **25**(1):44-52.

18. Sanderson S, Tatt ID, Higgins JP: **Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography**. *Int J Epidemiol* 2007, **36**(3):666-676.

19. Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, Devereaux PJ, Montori VM, Freyschuss B, Vist G *et al*: **GRADE guidelines 6. Rating the quality of evidence–imprecision**. *J Clin Epidemiol* 2011, **64**(12):1283-1293.

20. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y *et al*: **GRADE guidelines: 5. Rating the quality of evidence–publication bias**. *J Clin Epidemiol* 2011, **64**(12):1277-1282.

21. Saini P, Loke YK, Gamble C, Altman DG, Williamson PR, Kirkham JJ: **Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews**. *BMJ : British Medical Journal* 2014, **349**:g6501.

22. Rodgers MA, Brown JVE, Heirs MK, Higgins JPT, Mannion RJ, Simmonds MC, Stewart LA: **Reporting of industry funded study outcome data: comparison of confidential and published data on the safety and effectiveness of rhBMP-2 for spinal fusion**. *BMJ : British Medical Journal* 2013, **346**:f3981.

23. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V *et al*: **GRADE guidelines: 9. Rating up the quality of evidence**. *J Clin Epidemiol* 2011, **64**(12):1311-1316.

24. Bross ID: **Pertinency of an extraneous variable**. *Journal of chronic diseases* 1967, **20**(7):487-495.

25. Fewell Z, Davey Smith G, Sterne JAC: **The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study**. *American Journal of Epidemiology* 2007, **166**(6):646-655.

26. Kuss O, Miller M: **Unknown confounders did not bias the treatment effect when improving balance of known confounders in randomized trials**. *J Clin Epidemiol* 2020, **126**:9-16.

27. Glasziou P, Chalmers I, Rawlins M, McCulloch P: **When are randomised trials unnecessary? Picking signal from noise**. *BMJ* 2007, **334**(7589):349-351.

28. Sandberg L, Taavola H, Aoki Y, Chandler R, Norén GN: **Risk Factor Considerations in Statistical Signal Detection: Using Subgroup Disproportionality to Uncover Risk Groups for Adverse Drug Reactions in VigiBase**. *Drug Safety* 2020, **43**(10):999-1009.

29. Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, Alonso-Coello P, Glasziou P, Jaeschke R, Akl EA *et al*: **GRADE guidelines: 7. Rating the quality of evidence–inconsistency**. *J Clin Epidemiol* 2011, **64**(12):1294-1302.

30. Schandelmaier S, Briel M, Varadhan R, Schmid CH, Devasenapathy N, Hayward RA, Gagnier J, Borenstein M, van der Heijden GJMG, Dahabreh IJ *et al*: **Development of the Instrument to assess the Credibility of Effect Modification Analyses (ICEMAN) in randomized controlled trials and meta-analyses**. *Canadian Medical Association Journal* 2020, **192**(32):E901-E906.