GRADE approach to drawing conclusions from a network meta-analysis using a partially contextualised framework

Romina Brignardello-Petersen,[1] Ariel Izcovich,[2] Bram Rochwerg,[1] Ivan D Florez,[1,3] Glen Hazlewood,[4] Waleed Alhazanni,[1] Juan Yepes-Nuñez,[5] Nancy Santesso,[1] Gordon H Guyatt,[1] Holger J Schünemann,[1,6] on behalf of the GRADE working group

For numbered affiliations see end of the article.

Correspondence to:
R Brignardello-Petersen
brignarr@mcmaster.ca
(or @rominabrigpet on Twitter;
ORCID 0000-0002-6010-9900)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;371:m3907
http://dx.doi.org/10.1136/bmj.m3907

Accepted: 2 October 2020

This article describes GRADE (grading of recommendations assessment, development and evaluation) guidance on how to make conclusions from a network meta-analysis of interventions that includes individual randomised controlled trials for one outcome at a time. The guidance is based on a partially contextualised approach in which review authors must establish ranges of magnitudes of effect that represent a trivial to no effect, small but important effect, moderate effect, and large effect. The principles guiding this framework are that interventions should be grouped in categories, based on the magnitude of the effect; and that the judgments that place interventions in such categories should consider the estimates of effect, the certainty of the evidence, and the rankings. We describe and illustrate the four steps of this framework using an example.

SUMMARY POINTS

- Network meta-analysis (NMA) rarely establishes that, for a single outcome, one intervention is better than all others
- Classification of interventions into those with a trivial, small, moderate, or large effect can better reflect the results and is consistent with guidance by the GRADE (grading of recommendations assessment, development, and evaluation) working group on how to communicate findings
- This classification and the resulting conclusions should consider the estimates of effect, certainty of the evidence, and treatment rankings
- This article describes GRADE guidance on how to draw conclusions from NMA for one outcome using a partially contextualised approach, which categorises interventions according to the magnitude of effect and further considers the certainty of the evidence and rankings
- NMA users and reviewers that apply GRADE should use the new approach to ensure appropriate, informative, user friendly conclusions

Systematic review authors should draw conclusions on how interventions compare to others with regards to specific health outcomes considering the estimates of effects comparing those interventions, and the certainty of the evidence (confidence in evidence, quality of evidence).[1] When review authors conduct network meta-analysis (NMA), they might also have information on how likely each intervention is the most beneficial or harmful for the outcome (rankings). The large amount of information that emerges from an NMA—that is, a relative estimate and its certainty for each comparison, in addition to the rankings—raises challenges in reaching appropriate conclusions that consider all key information.

The GRADE (grading of recommendations assessment, development, and evaluation) working group has presented guidance for evaluating the certainty of the evidence in NMA,[2 3] how to avoid spurious judgments when addressing imprecision,[4] and how to assess incoherence.[5] In addition, we have provided suggestions for how to present the findings from an NMA in a summary of findings table.[6] No guidance so far, however, exists on how to draw conclusions from the comparisons in the NMA.

Based on our experience (and that of the experts who provided feedback), it is unlikely that one intervention is definitely superior to all other interventions for a particular outcome—which is especially the case in large networks, for several reasons. Firstly, although treatments can be ranked statistically from the best to the worse, the effects of interventions that rank higher might not be importantly different than those of interventions that rank lower. In other words, differences in an effect might be trivial, small but important, moderate, or large, and the implications vary importantly across these categories.[7] Moreover, certainty of the evidence usually varies from high to very low across the often many comparisons in an NMA. Interventions that rank high might have low or very low certainty evidence, while other interventions might rank low and have higher associated certainty.[8 9] Therefore, there is seldom one intervention with high or moderate certainty evidence indicating that it is clearly superior compared with all other interventions.

This article describes how to interpret findings of an NMA for each outcome. Depending on the context, interpretation can be done with a minimally contextualised framework, in which value judgments are made regarding the importance of the magnitude of the effects are minimised; or a partially contextualised framework, in which review authors consider the importance of the magnitude of the effects on an

outcome without regard of other outcomes. This article focuses on the partially contextualised framework.

The description of this framework assumes familiarity with the basic concepts of NMA, the implications of GRADE's certainty of the evidence, and the degress of contextualisation. This article constitutes official guidance from the GRADE working group. This framework was developed, tested, and refined by the named authors with feedback from the entire GRADE working group that ultimately approved the paper as GRADE guidance.

Methods

This project was conducted under the auspices of the GRADE NMA project group. First, we conducted a systematic survey of the literature showing that no methods have been proposed to make conclusions from an NMA for one outcome that simultaneously considers the results from an NMA and the certainty in the evidence. A core team of experts in systematic review methodology and NMAs then developed an initial framework using a minimally contextualised framework.[10]

Reviewing the potential benefits of contextualisation, another team of experts (HJS, NS, RB-P) proposed an alternative framework in which the magnitude of the effect and its healthcare interpretation has a central role. This contextualised framework is built on GRADE's Evidence to Decision frameworks[11-13] and GRADE's guidance on how to interpret findings from pairwise comparisons.[14]

We obtained feedback about this initial framework from other experts in systematic reviews methodology, biostatisticians, and systematic review authors, both with and without experience in NMA, and who were and were not members of the GRADE working group. We also tested this framework in several examples (some examples included in the appendix). Finally, we presented the final framework to the GRADE working group at the meeting in Hamilton, Canada (June 2019) and Adelaide, Australia (November 2019), to obtain approval to publish this framework as GRADE guidance.

Results

The partially contextualised framework to make conclusions from NMA has two guiding principles and four steps, which we describe below. The principles are similar as those for the minimally contextualised framework, but the conceptual underpinnings, some steps, and judgments required differ substantially.

Guiding principles

The framework to draw conclusions from NMA, for one outcome, is based on two principles. Firstly, categories of interventions should be considered (eg, those with a trivial effect, small effect, moderate effect, or large effect). The effect can be either a benefit or a harm, depending on the context. In addition, depending on the results of each NMA, there might not be interventions in all the categories that describe the magnitude of the effect. Secondly, the judgments that place interventions in categories will rely on the estimates of effect, and the intervention rankings; and the conclusions will then consider the certainty of the evidence. None of the pieces of information can be used alone to determine whether an intervention is better than others.

Use of partially contextualised framework to draw conclusions from network meta-analyses

The process for drawing conclusions from NMAs has four steps. Review authors must conduct this process after they have finalised ratings of the certainty of the evidence for each comparison in the NMA. We illustrate each of the steps using an example NMA of pharmacological and nutritional interventions for treating acute diarrhoea and gastroenteritis in children.[15] The primary outcome of this systematic review was diarrhoea duration, and the treatment effects were measured as difference in hours. The NMA included 138 randomised controlled trials in which researchers recruited 20 256 participants and assessed the effects of 27 interventions. The network has a complex geometry (fig 1), with 62 direct comparisons and 289 indirect comparisons. We present more examples in the appendix,[16-18] which include dichotomous outcomes.

Step 1: Choose reference intervention and thresholds for effects

Review authors should choose the intervention most connected to the other interventions in the network and use that intervention as a reference. Network estimates that are calculated with direct evidence are more likely to be judged as higher certainty evidence than those calculated with indirect evidence only, which results in classifying the treatments using the highest certainty evidence. In addition, this increases the likelihood of better differentiating between the interventions and achieving a more informative classification than if the classification was based on lower certainty evidence.

The reference intervention must be used for the process of drawing conclusions, but it does not necessarily have to be used as the reference for the purpose of presenting results if other treatments less connected to the network are more clinically meaningful as a reference. In the NMA of interventions for acute diarrhoea in children,[15] the reference intervention was standard treatment that included arms characterised as "no active treatment," "placebo," or "only oral rehydration solution."

Similar to GRADE guidance for communicating the results from systematic reviews,[14] reviewers assessing the evidence must make judgments for what constitutes a trivial to no effect, small but important effect, moderate effect, and large effect. These judgments will serve as the basis for the classification of the interventions into groups, and should be established by informed review teams that possess the required health knowledge, ideally based on input from key stakeholders. The process for making

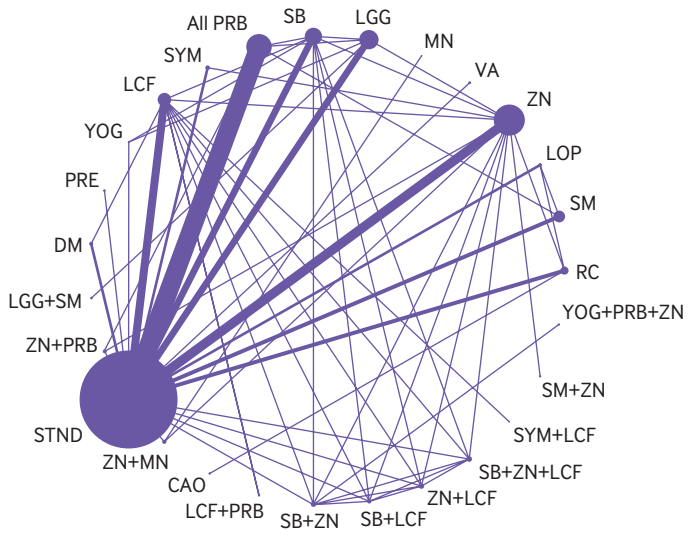

Fig 1 | Network plot of interventions for treating acute diarrhoea and gastroenteritis in children, for the outcome diarrhoea duration.[15] PRB=all probiotics; SB=*Saccharomyces boulardii*; LGG=*Lactobacillous rhamnosus GG*; MN=micronutrients; VA=vitamin A; ZN=zinc; LOP=loperamide; SM=smectite; RC=racecadotril; YOG=yoghurt; SYM=symbiotics; LCF=lactose free formula; CAO=Kaolin-pectin; STND=placebo or standard care; DM=diluted milk; PRE=prebiotics

these choices should be explicit and transparent; they might not be the same, even within the same NMA in different contexts. In addition, and consistent with GRADE guidance, these choices should be made based on absolute estimates rather than relative estimates of effect.

Absolute values (as in our example here) will be the natural report for continuous outcomes. The same, however, is not true for binary outcomes in which the NMA will yield estimates of relative effect that then need translation into absolute effect. Translation to absolute effects is necessary because judgments of importance (or judgments of magnitude of effect as small, moderate, or large) cannot be made on the basis of relative effects. For example, a 50% relative reduction with a baseline risk of 2% represents a 1% absolute risk reduction that might be considered unimportant, and if important as a small effect. That same 50% relative risk reduction, in the setting of a baseline risk of 40%, represents a 20% absolute risk reduction that could be judged as very important and large. For the purpose of this illustration, the authors of the review of interventions for acute diarrhoea[15] determined that a small but important effect was a reduction or increase in diarrhoea duration from 3 to 12 hours, a moderate effect was a reduction or increase from 12 to 24 hours, and a large effect was a reduction or increase of 24 hours or more (fig 2).

Step 2: Classification based on comparison with reference

In this step, review authors should use the point estimate comparing each of the interventions against the reference. This point estimate, which represents the best estimate of effect, should be assessed against the thresholds for small, moderate, and large effects established in the previous step. Depending on the point estimate, each intervention should be classified as being in the range of trivial, small but important, moderate, or lage effects. Depending on its direction, the effect can either be a benefit or a harm when compared with the reference. Figure 2 illustrates this classification in the NMA of interventions for acute diarrhoea.[15]

The number of groups that result from this classification will depend on the specific NMA. The NMA of interventions for acute diarrhoea in children[15] had five groups of interventions: small harm, trivial to no effect, small benefit, moderate benefit, or large benefit (table 1).

Step 3: Identification according to certainty of evidence

In this third step, review authors should use the certainty of the evidence for every treatment, when compared with the reference, in order to make the level of certainty explicit for each comparison with the reference. Review authors can choose to group interventions with high or moderate certainty evidence together, and those with low or very low certainty evidence. This classification might be reasonable in a network with several interventions and with many comparisons across all levels of evidence; however, interventions with low or very low certainty evidence should not be grouped together if most of the interventions have either low or very low certainty when compared with the reference, because review authors would lose the opportunity to differentiate according to evidence certainty. Table 2 shows the classification of interventions for acute diarrhoea and gastroenteritis in children,[15] sorted by groups according to the magnitude of the effect and specifying the certainty of the evidence.

Review authors should draw conclusions about how likely each intervention has the magnitude of effect specified according to the certainty of the evidence. For instance, authors can state that "LGG [*Lactobacillous rhamnosus GG*] probably has moderate benefits when compared to standard therapy," and that "Micronutrients may have trivial to no effect compared to standard therapy."[14]

Step 4: Checking consistency with pairwise comparisons and rankings

In the fourth step, review authors should make sure that the classification is consistent with the pairwise comparisons not considered in the process (that is, the comparisons between pairs of interventions that are not the reference) and their certainty. The classification can be reviewed and adjusted if the pairwise comparisons suggest a different conclusion with high or moderate certainty evidence.

In this step, reviewers should consider the possibility that an intervention appears superior to another in relation to the reference intervention but not in a direct comparison between the two. For instance,

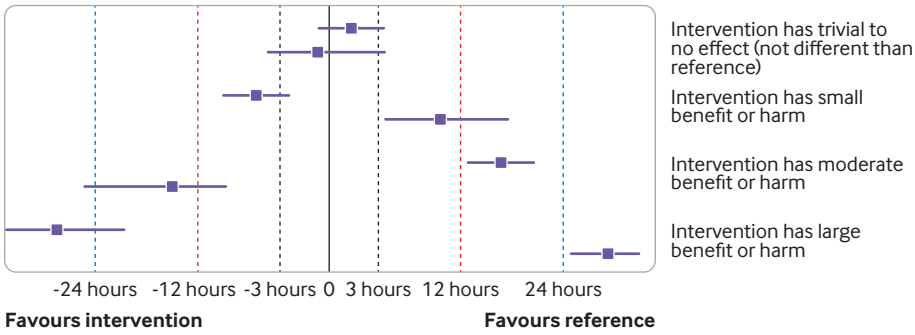

Fig 2 | Example classification based on comparison with the reference intervention. Thresholds correspond to those of the network meta-analysis of interventions for acute diarrhoea in children and diarrhoea duration.[15] This classification is step 2 of the partially contextualised framework process and has not yet incorporated the certainty of the evidence, which should be considered in the final conclusions

consider a situation in which intervention A achieves a large benefit relative to placebo (the reference) and intervention B achieves only a moderate benefit relative to placebo. Intervention A will then be ranked higher than intervention B, but this ranking would be problematic if the interventions are directly compared with each other and B does better than A in achieving benefit. Although unlikely to happen, reviewers should be alert to these situations.

When looking at the indirect comparisons between non-reference interventions in the NMA of interventions for acute diarrhoea in children,[15] we saw no indications that the classification was not appropriate. For example, when looking at the comparison between *Saccharomyces boulardii* + zinc (classified as moderate certainty of a large beneficial effect) and yoghurt (classified as very low certainty of a moderate beneficial effect), the estimate comparing them was a mean difference of −22.96 hours of diarrhoea duration (95% confidence interval −42.15 to −4.44, very low quality evidence). This difference suggests that *S boulardii* + zinc could have a larger benefit than yoghurt. Similarly, when comparing interventions smectite + zinc (classified as moderate certainty of a large benefit) with vitamin A (classified as very low certainty of a small benefit), the estimate (mean difference −29.54 hours (−56.09 to −2.84), moderate quality evidence) suggests that smectite + zinc (M) could have a larger benefit than vitamin A.

Review authors can use also the rankings, rank probabilities, SUCRA (surface under the cumulative ranking curve) values, or P scores, if available, to check whether the classification in the groups is sensible, and can adjust the classification if necessary. For example, consider again an intervention with a large effect ranked higher than an intervention with a moderate effect; if the first intervention has a considerably lower SUCRA value than the second intervention, it suggests a problem. In the NMA of interventions for acute diarrhoea in children,[15] the SUCRA values decreased from the intervention group with a large benefit to the intervention group with a large harm (table 2), indicating no need to revise the classification.

If the assumptions of NMA are met, the likelihood that step 4 changes the classification is low. Review authors should consider the amount of information provided by the pairwise comparisons not considered in previous steps, and safeguard against any possible mistake. After finishing these four steps, review authors can describe this classification to make their conclusions. According to GRADE guidance on how to communicate findings, the conclusions in this example[15] are:

- When considering all the interventions, symbiotics have a large beneficial effect on diarrhoea duration
- When considering all the interventions, *S boulardii* + zinc and smectite + zinc probably have a large beneficial effect on diarrhoea duration
- When considering all the interventions, zinc + probiotics might have a large beneficial effect on diarrhoea duration
- When considering all the interventions, zinc + lactose-free formula, zinc, loperamide, and zinc + micronutrients probably have a moderate beneficial effect on diarrhoea duration
- When considering all the interventions, all probiotics, racecadotril, *S boulardii*, and *S boulardii* + zinc + lactose-free formula classified as low certainty of a moderate beneficial effect might have a moderate beneficial effect on diarrhoea duration
- When considering all the interventions, micronutrients might have a trivial effect on diarrhoea duration
- For the rest of the interventions, the effect is uncertain because the certainty on the evidence was very low.

Discussion

This article describes the GRADE working group guidance for drawing conclusions from an NMA using a partially contextualised framework. This framework allows review authors to classify interventions in different groups considering the magnitude of effect, certainty of the evidence, and rankings, if available; and to draw appropriate conclusions. The number

Table 1 | Classification of each intervention for acute diarrhoea when compared with standard treatment, based on example network meta-analysis of interventions for acute diarrhoea in children[15]

Intervention	Effect on diarrhoea duration (hours; mean difference (95%CI))	Classification of intervention
All probiotics	−19.36 (−23.66 to −15.09)	Moderate beneficial effect
Diluted milk	3.02 (−14.32 to 8.41)	Small harmfull effect
Kaolin-pectin	−5.32 (−33.76 to 22.83)	Small beneficial effect
Lactose-free formula	−12.50 (−19.04 to −5.99)	Moderate beneficial effect
Lactose-free formula + probiotics	−13.27 (−35.96 to 9.19)	Moderate beneficial effect
Lactobacillous rhamnosus GG	−22.74 (−28.81 to −16.68)	Moderate beneficial effect
L rhamnosus GG + smectite	−51.08 (−64.30 to −37.85)	Large beneficial effect
Loperamide	−17.79; (−30.35 to −5.65)	Moderate beneficial effect
Micronutrients	−0.68 (−33.29 to 32.79)	Little to no effect
Prebiotics	−15.62 (−42.42 to 11.28)	Moderate beneficial effect
Racecadotril	−17.19 (−24.65 to −9.76)	Moderate beneficial effect
*Saccharomyces boulardi*i	−16.48 (−23.33 to −9.69)	Moderate beneficial effect
S boulardii + lactose-free formula	−12.32 (−30.01 to 5.98)	Moderate beneficial effect
S boulardii + zinc	−39.45 (−52.45 to −26.73)	Large beneficial effect
S. boulardii + zinc + lactose-free formula	−16.74 (−36.05 to 2.72)	Moderate beneficial effect
Smectite	−23.90 (−30.80 to −16.96)	Moderate beneficial effect
Smectite + zinc	−35.63 (−57.57 to −13.16)	Large beneficial effect
Symbiotics	−26.26 (−36.14 to −16.22)	Large beneficial effect
Symbiotics + lactose-free formula	−32.11 (−53.01 to −11.33)	Large beneficial effect
Vitamin A	−5.95 (−21.43 to 9.32)	Small beneficial effect
Yoghurt	−16.43 (−30.49 to −2.05)	Moderate beneficial effect
Yoghurt + probiotics + zinc	−15.63 (−56.82 to 26.63)	Moderate beneficial effect
Zinc	−18.38 (−23.39 to −13.45)	Moderate beneficial effect
Zinc + lactose-free formula	−21.37 (−36.54 to −6.13)	Moderate beneficial effect
Zinc + micronutrients	−17.76 (−31.77 to −4.13)	Moderate beneficial effect
Zinc + probiotics	−29.39 (−40.26 to −18.57)	Large beneficial effect

of resulting categories depends on the evidence available, how many interventions are included in the NMA, how the interventions compare with one another, and the thresholds of magnitude of effect. This framework follows similar guiding principles to the minimally contextualised framework[10] with one important difference.

The main difference between the minimally contextualised framework and the partially contextualised framework is that the categorisation in the partially contextualised framework does not emphasise imprecision over other GRADE domains to determine whether an effect is present (imprecision together with all other GRADE domains is considered when rating the certainty of the network estimate). In contrast, in the minimally contextualised framework we present elsewhere,[10] the initial classification relative to the reference standard focuses (as does the subsequent classification considering differences between non-reference interventions) on whether the confidence interval excludes an established threshold.

Using the minimally contextualised framework, for instance, in comparison to the reference standard and using a no-effect decision threshold, categorisation would differ for an absolute risk reduction of 20% (95% confidence interval 1% to 39%) versus the same point estimate with a 95% confidence interval of −1% to 41%. Using the partially contextualised approach, the initial classification would be made on the basis of the point estimate (and with the same point estimate relative to the reference would be placed in the same category), and whether the confidence interval crosses the null would be irrelevant. This partially contextualised approach acknowledges, for example, that an intervention effect with a confidence interval of 1% to 39% that is rated down for risk of bias should not be more trustworthy than an effect with a confidence interval of −1% to 41% that is rated down for imprecision.

However, both frameworks, and indeed almost any system of classification, are vulnerable to the arbitrariness of thresholds. In the minimally contextualised framework, one threshold of focus is no difference between interventions. In the partially contextualised framework, the threshold of focus is the boundaries between ranges: in the current NMA example,[15] a difference of 3.01 hours would be classified differently from a difference of 2.99 hours.

The partially contextualised framework might be particularly appealing in contexts where the specific magnitude of the potential benefit or harm (and whether it represents a trivial, small, moderate, or large effect) are key in helping review authors draw conclusions. This categorisation has an important role in the development of healthcare guidelines when panels judge the balance between health benefits and harms. In such contexts, this framework allows contextualisation through the thresholds of small, moderate, and large effects and other Evidence to Decision criteria.

This framework is described as partially contextualised because it requires the reviewer of the evidence to make explicit and transparent value judgments regarding magnitudes of effect that represent small, moderate, or large benefits or harms. Review authors make value judgments, regardless

Table 2 | Classification of interventions based on network meta-analysis of interventions for acute diarrhoea and gastroenteritis in children[15]

Classification* of intervention	Intervention	Effect on diarrhoea duration (hours; mean difference (95% CI))	Surface under the cumulative ranking curve (95% CI)	Certainty of evidence
Large beneficial effect	Lactose-free formula + smectite	−51.08 (−64.30 to −37.85)	1.00 (0.92 to 1.00)	Very low
	Saccharomyces boulardii + zinc†	−39.45 (−52.45 to −26.73)†	0.92 (0.77 to 1.00)†	Moderate†
	Smectite + zinc†	−35.63 (−57.57 to −13.16)†	0.88 (0.35 to 1.00)†	Moderate†
	Symbiotics + lactose-free formula	−32.11 (−53.01 to −11.33)	0.85 (0.27 to 1.00)	Very low
	Zinc + probiotics	−29.39 (−40.26 to −18.57)	0.81 (0.5 to 0.96)	Low
	Symbiotics†	−26.26 (−36.14 to −16.22)†	0.77 (0.38 to 0.92)†	High†
Moderate beneficial effect	Smectite	−23.90 (−30.80 to −16.96)	0.69 (0.42 to 0.88)	Very low
	Lactobacillous rhamnosus GG	−22.74 (−28.81 to −16.68)	0.65 (0.38 to 0.85)	Low
	Zinc + lactose-free formula†	−21.37 (−36.54 to −6.13)†	0.61 (0.19 to 0.92)†	Moderate†
	All probiotics	−19.36 (−23.66 to −15.09)	0.54 (0.31 to 0.73)	Low
	Zinc†	−18.38 (−23.39 to −13.45)†	0.50 (0.27 to 0.69)†	Moderate†
	Loperamide†	−17.79; (−30.35 to −5.65)†	0.46 (0.15 to 0.85)†	Moderate†
	Zinc + micronutrients†	−17.76 (−31.77 to −4.13)†	0.46 (0.15 to 0.85)†	Moderate†
	Racecadotril	−17.19 (−24.65 to −9.76)	0.46 (0.23 to 0.73)	Low
	S boulardii + zinc + lactose-free formula	−16.74 (−36.05 to 2.72)	0.42 (0.08 to 0.88)	Low
	S boulardii	−16.48 (−23.33 to −9.69)	0.42 (0.19 to 0.69)	Low
	Yoghurt	−16.43 (−30.49 to −2.05)	0.42 (0.11 to 0.85)	Very low
	Yoghurt + probiotics + zinc	−15.63 (−56.82 to 26.63)	0.38 (0.00 to 1.00)	Very low
	Prebiotics	−15.62 (−42.42 to 11.28)	0.38 (0.00 to 0.96)	Very low
	Lactose-free formula + probiotics	−13.27 (−35.96 to 9.19)	0.31 (0.00 to 0.88)	Very low
	Lactose-free formula	−12.50 (−19.04 to −5.99)	0.31 (0.15 to 0.54)	Very low
	S boulardii + lactose-free formula	−12.32 (−30.01 to 5.98)	0.27 (0.04 to 0.81)	Very low
Small beneficial effect	Vitamin A	−5.95 (−21.43 to 9.32)	0.19 (0.00 to 0.61)	Very low
	Kaolin-pectin	−5.32 (−33.76 to 22.83)	0.15 (0.00 to 0.89)	Very low
Trivial to no effect (not different than placebo)	Micronutrients	−0.68 (−33.29 to 32.79)	0.08 (0.00 to 0.85)	Low
Small harmful effect	Diluted milk	3.02 (−14.32 to 8.41)	0.04 (0.00 to 0.23)	Very low

*A suggested format of presentation could include different colours or shades according to the magnitude of effect; this presentation format has not been user tested and is not guidance from the GRADE working group.
†Presence of high or moderate certainty evidence.

of the degree of contextualisation, by identifying critical and important outcomes for inclusion in their systematic review. Currently, many systematic reviews are done for specific purposes, for example, to inform a guideline or a health technology assessment. Thus, the guideline panel will require value judgments to make recommendations.

We have not established rules of thumb for these judgments, and they might vary across different settings. In the context of systematic reviews that are designed to inform guidelines, these judgments should be made by the panel of experts and should be informed by evidence regarding patients' values regarding each outcome.[19 20] Ideally, the judgments are made by establishing close collaboration between the review team and members of the decision making group (eg, the guideline panel) early on in the process of developing recommendations.[21] Authors of guidelines using existing systematic reviews can establish their own thresholds and reclassify the interventions according to their needs that, if made transparent, are then reviewed and modified by descision makers.[22] In the context of systematic reviews that are not specifically designed to inform guidelines, these judgments can be made by the clinical experts involved in the systematic review team, considering the relative importance of each outcome.

We developed this framework after we recognised that contextualising the classification by considering the importance of the magnitude of the effect could be desirable in many instances. In NMAs where the evidence for most of the comparisons is indirect, and that are more likely to have wide and imprecise estimates, use of this partially contextualised framework also maximises the chances to differentiate among interventions. In this partially contextualised framework, the width of the confidence intervals is accounted for when assessing imprecision and not used again for drawing conclusions.

The main limitation of this framework is that the conclusions depend substantially on the thresholds established, but from our experience in working with many guideline panels we are confident that calibration takes place easily. Thus, while this limitation might be considered a problem, it is no different from what happens when systematic reviewer authors draw conclusions regarding the magnitude of an effect in the context of any meta-analysis. When using this approach, however, review authors must be explicit about the thresholds and should establish them using absolute estimates of effect. Thus, the step of establishing the thresholds is likely to make review authors more aware of the implication of such thresholds and to make them put more thought into the thresholds than usual.

Secondly, although use of only one intervention as the reference might mean that review authors can ignore a large amount of information, the fourth step of the process requires review authors to confirm that the pairwise comparisons between non-reference interventions and the rankings are consistent with the classification. Therefore, review authors have the

chance to adjust the classification using the information not considered initially (although this adjustment is probably not needed in an NMA that was designed appropriately and meets the basic assumptions of an NMA). Finally, despite some concern about use of the point estimates alone for making conclusions, the point estimate has been argued to be the best estimate of effect and information regarding any uncertainty reflected in confidence intervals is captured in the rating of the certainty of the evidence.

In summary, this partially contextualised framework guides review authors to make conclusions from NMA, considering all the crucial pieces of information. This framework is likely to be the most appropriate in scenarios where most of the evidence is indirect and when the systematic reviews with NMAs are conducted to inform decisions such as in guidelines or coverage decisions following an health technology assessment.

Author affiliations

[1]Department of Health Research Methods, Evidence and Impact, McMaster University, 1280 Main Street West, Hamilton, ON L8S 4L8, Canada

[2]Internal Medicine Service, German Hospital, Buenos Aires, Argentina

[3]Department of Pediatrics, School of Medicine, University of Antioquia, Medellín, Colombia

[4]Department of Medicine, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada

[5]School of Medicine, Universidad de los Andes, Bogota, Colombia

[6]GRADE Centre and Department of Medicine, McMaster University, Hamilton, ON, Canada

We thank all members of the GRADE NMA project group and GRADE working group for their input on this manuscript, in particular to Monica Hultcranz, Reem Mustafa, Derek Chu, and Ilse Verst.

Contributors: RB-P, JY-N, and HJS developed the principles and initial version of the framework. BR, NS, and GHG provided input that resulted important modifications. RB-P and AI tested the framework in several examples. IDF, BR, GH, and WA provided data from the examples included in this article. RB-P, AI, and HJS drafted and edited the manuscript, based on feedback from all the authors and members of the GRADE working group. All authors approved the final version of the manuscript. HJS, who had a major role at all stages of this project, is the guarantor of this article. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: This project was did not receive funding.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work

Ethical approval: Not applicable. All the work was developed using published data.

The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Provenance and peer review: Not commissioned; externally peer reviewed.

Patient and public involvement: Due to the nature of this work, we did not include patients and public.

1 Schünemann HJ, Higgins JPT, Vist GE, et al. Completing 'Summary of findings' tables and grading the certainty of the evidence. In: Higgins JPT, Thomas J, Chandler J, et al. eds. *Cochrane Handbook for Systematic Reviews of Interventions*. 2nd ed. John Wiley & Sons, 2019. doi:10.1002/9781119536604.ch14

2 Brignardello-Petersen R, Bonner A, Alexander PE, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. *J Clin Epidemiol* 2018;93:36-44. doi:10.1016/j.jclinepi.2017.10.005

3 Puhan MA, Schünemann HJ, Murad MH, et al, GRADE Working Group. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. *BMJ* 2014;349:g5630. doi:10.1136/bmj.g5630

4 Brignardello-Petersen R, Murad MH, Walter SD, et al, GRADE Working Group. GRADE approach to rate the certainty from a network meta-analysis: avoiding spurious judgments of imprecision in sparse networks. *J Clin Epidemiol* 2019;105:60-7. doi:10.1016/j.jclinepi.2018.08.022

5 Brignardello-Petersen R, Mustafa RA, Siemieniuk RAC, et al, GRADE Working Group. GRADE approach to rate the certainty from a network meta-analysis: addressing incoherence. *J Clin Epidemiol* 2019;108:77-85. doi:10.1016/j.jclinepi.2017.10.005

6 Yepes-Nunez JJ, Li SA, Guyatt G, et al. Development of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) summary of findings (SoF) table for network meta-analysis. *J Clin Epidemiol* 2019;2:30317-2. doi:10.1016/j.jclinepi.2019.04.018

7 Schünemann HJ, Wiercioch W, Brozek J, et al. GRADE Evidence to Decision (EtD) frameworks for adoption, adaptation, and de novo development of trustworthy recommendations: GRADE-ADOLOPMENT. *J Clin Epidemiol* 2017;81:101-10. doi:10.1016/j.jclinepi.2016.09.009

8 Brignardello-Petersen R, Guyatt GH. β-Blockers in heart failure--are all created equal?*Pol Arch Med Wewn* 2013;123:204-5. doi:10.20452/pamw.1720

9 Brignardello-Petersen R, Rochwerg B, Guyatt GH. What is a network meta-analysis and how can we use it to inform clinical practice?. *Pol Arch Med Wewn* 2014;124:659-60. doi:10.20452/pamw.2546

10 Brignardello-Petersen R, Florez ID, Izcovich A, et al, on behalf of the GRADE working group. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. *BMJ* 2020;371:m3900.

11 Alonso-Coello P, Oxman AD, Moberg J, et al, GRADE Working Group. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. *BMJ* 2016;353:i2089. doi:10.1136/bmj.i2089

12 Alonso-Coello P, Schünemann HJ, Moberg J, et al, GRADE Working Group. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ* 2016;353:i2016. doi:10.1136/bmj.i2016

13 Schünemann HJ, Mustafa R, Brozek J, et al, GRADE Working Group. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. *J Clin Epidemiol* 2016;76:89-98. doi:10.1016/j.jclinepi.2016.01.032

14 Santesso N, Glenton C, Dahm P, et al, GRADE Working Group. GRADE guidelines 26: informative statements to communicate the findings of systematic reviews of interventions. *J Clin Epidemiol* 2020;119:126-35. doi:10.1016/j.jclinepi.2019.10.014

15 Florez ID, Veroniki AA, Al Khalifah R, et al. Comparative effectiveness and safety of interventions for acute diarrhea and gastroenteritis in children: A systematic review and network meta-analysis. *PLoS One* 2018;13:e0207701. doi:10.1371/journal.pone.0207701

16 Hazlewood GS, Barnabe C, Tomlinson G, Marshall D, Devoe D, Bombardier C. Methotrexate monotherapy and methotrexate combination therapy with traditional and biologic disease modifying antirheumatic drugs for rheumatoid arthritis: abridged Cochrane systematic review and network meta-analysis. *BMJ* 2016;353:i1777. doi:10.1136/bmj.i1777

17 Alhazzani W, Alshamsi F, Belley-Cote E, et al. Efficacy and safety of stress ulcer prophylaxis in critically ill patients: a network meta-analysis of randomized trials [A: correction in: *Intensive Care Med* 2018;44:277-78]. *Intensive Care Med* 2018;44:1-11. doi:10.1007/s00134-017-5005-8

18 Rochwerg B, Neupane B, Zhang Y, et al. Treatment of idiopathic pulmonary fibrosis: a network meta-analysis. *BMC Med* 2016;14:18. doi:10.1186/s12916-016-0558-x

19 Zhang Y, Alonso-Coello P, Guyatt GH, et al. GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences-Risk of bias and indirectness. *J Clin Epidemiol* 2019;111:94-104. doi:10.1016/j.jclinepi.2018.01.013

20 Zhang Y, Coello PA, Guyatt GH, et al. GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences-inconsistency, imprecision, and other domains. *J Clin Epidemiol* 2019;111:83-93. doi:10.1016/j.jclinepi.2018.05.011

21 Schünemann HJ, Lerda D, Dimitrova N, et al, European Commission Initiative on Breast Cancer Contributor Group. Methods for development of the European Commission Initiative on Breast Cancer Guidelines: recommendations in the era of guideline transparency. *Ann Intern Med* 2019;171:273-80. doi:10.7326/M18-3445

22 Schünemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision?/ *Clin Epidemiol* 2016;75:6-15. doi:10.1016/j.jclinepi.2016.03.018

Web appendix: Supplementary material