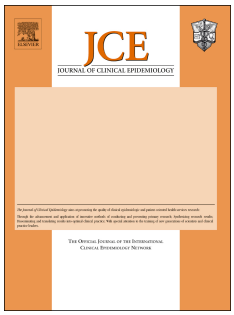


# GRADE guidelines 26: Informative statements to communicate the findings of systematic reviews of interventions

Nancy Santesso, Claire Glenton, Philipp Dahm, Paul Garner, Elie Akl, Brian Alper, Romina Brignardello-Petersen, Alonso Carrasco-Labra, Hans De Beer, Monica Hultcrantz, Ton Kuijpers, Joerg Meerpohl, Rebecca Morgan, Reem Mustafa, Nicole Skoetz, Shahnaz Sultan, Charles Wiysonge, Gordon Guyatt, Holger J. Schünemann, for the GRADE Working Group

PII: S0895-4356(19)30416-0

DOI: https://doi.org/10.1016/j.jclinepi.2019.10.014

Reference: JCE 10008

To appear in: *Journal of Clinical Epidemiology*

Received Date: 17 May 2019

Revised Date: 24 August 2019

Accepted Date: 7 October 2019

Please cite this article as: Santesso N, Glenton C, Dahm P, Garner P, Akl E, Alper B, Brignardello-Petersen R, Carrasco-Labra A, De Beer H, Hultcrantz M, Kuijpers T, Meerpohl J, Morgan R, Mustafa R, Skoetz N, Sultan S, Wiysonge C, Guyatt G, Schünemann HJ, for the GRADE Working Group, GRADE guidelines 26: Informative statements to communicate the findings of systematic reviews of interventions, *Journal of Clinical Epidemiology* (2019), doi: https://doi.org/10.1016/j.jclinepi.2019.10.014.

# GRADE guidelines 26: Informative statements to communicate the findings of systematic reviews of interventions

Nancy Santesso[1], Claire Glenton[2], Philipp Dahm[3], Paul Garner[4], Elie Akl[5], Brian Alper[6], Romina Brignardello-Petersen[1], Alonso Carrasco-Labra[1], Hans De Beer[7], Monica Hultcrantz[8], Ton Kuijpers[9], Joerg Meerpohl[10], Rebecca Morgan[1], Reem Mustafa[1,11], Nicole Skoetz[12], Shahnaz Sultan[13], Charles Wiysonge[14], Gordon Guyatt[1,15], Holger J. Schünemann[1,15] for the GRADE Working Group

[1]Department of Health Research Methods, Evidence and Impact, Cochrane Canada, MacGRADE Centre and Michael G. DeGroote Cochrane Canada Centre, McMaster University, 1280 Main St East, Hamilton, Canada, L8S 4L8
santesna@mcmaster.ca, alonsocarrascolabra@gmail.com, brignarr@mcmaster.ca, morganrl@mcmaster.ca, guyatt@mcmaster.ca, Holger.schunemann@mcmaster.ca

[2]Cochrane Norway and the Informed Health Choices Research Centre, Norwegian Institute of Public Health, Postboks 222 Skøyen, Sandakerveien 24C, inngang D11, 0213, Oslo, Norway.
Claire.Glenton@fhi.no

[3]Minneapolis VA Health Care System, Urology Section 112D, One Veterans Drive, Minneapolis, MN 55417
pdahm@umn.edu

[4]Centre for Evidence Synthesis in Global Health, Liverpool School of Tropical Medicine, Liverpool, United Kingdom
Paul.Garner@lstmed.ac.uk

[5]Department of Internal Medicine, American University of Beirut, Lebanon, P.O.Box 11-0236
ea32@aub.edu.lb

[6]EBSCO Health, Innovations and Evidence-Based Medicine Development, 10 Estes Street, Ipswich, MA, USA 01938, and University of Missouri-Columbia, Department of Family and Community Medicine, Columbia, MO, USA
BAlper@ebsco.com

[7]Guide2Guidance, Lemelerberg 7, 3524 LC Utrecht, The Netherlands
jjadebeer@ziggo.nl

[8]Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU), S:t Eriksgatan 117, SE-102 33, Stockholm, Sweden
monica.hultcrantz@sbu.se

[9]Dutch College of General Practitioners (NHG), Department of guideline development and research, Mercatorlaan 1200, 3528 BL Utrecht, The Netherlands
T.Kuijpers@nhg.org

[10]Institute for Evidence in Medicine, Breisacher Strasse 153, 79110 Freiburg, Germany
meerpohl@ifem.uni-freiburg.de

[11]Department of Internal Medicine, Division of Nephrology and Hypertension, University of Kansas Medical Center, 3901 Rainbow Blvd, MS3002, Kansas City, KS 66160, United States
ramustafa@gmail.com

[12]University of Cologne, Faculty of Medicine and University Hospital Cologne, Department I of Internal Medicine, Kerpener Str. 62, 50931 Cologne, Germany.
nicole.skoetz@uk-koeln.de

[13]Division of Gastroenterology and Hepatology, and Nutrition, University of Minnesota, Minneapolis Veterans Affairs Healthcare System, Minneapolis, MN, USA
516 Delaware St. SE, 1st Floor, Phillips-Wangsteen Building, MMC 36, Minneapolis, MN 55455
shazi0314@gmail.com

[14]Cochrane South Africa, South African Medical Research Council, Cape Town, South Africa; School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa; and Department of Global Health, Stellenbosch University, Cape Town, South Africa.
Cochrane South Africa, South African Medical Research Council, Francie van Zijl Drive, Parow Valley, 7501, Cape Town, South Africa
charles.wiysonge@MRC.AC.ZA

[15]Department of Medicine, McMaster University, 1280 Main St East, Hamilton, Canada, L8S 4L8

**GRADE Working Group**

**Declarations of interest:** All authors confirm they have no direct financial conflicts of interest.

**Authorship**

Nancy Santesso: Conceptualization, design, data acquisition, analysis, interpretation, writing original draft and revisions. Gordon Guyatt and Holger Schünemann: conceptualization and design, interpretation, critical revisions. Claire Glenton, Philipp Dahm, Paul Garner, Elie Akl, Brian Alper, Romina Brignardello-Petersen, Alonso Carrasco-Labra, Hans De Beer, Monica Hultcrantz, Ton Kuijpers, Joerg Meerpohl, Rebecca Morgan[1], Reem Mustafa, Nicole Skoetz, Shahnaz Sultan, Charles Wiysonge: conceptualization, interpretation, critical revisions. All authors approved the final version. Members of the GRADE Working Group contributed to the conceptualization and interpretation.

**Funding**
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Abstract** (200 words)

**Objectives:** Clear communication of systematic review findings will help readers and decision makers. We built on previous work to develop an approach that improves the clarity of statements to convey findings and that draws on Grading of Recommendations Assessment, Development and Evaluation (GRADE).

**Study Design and Setting:** We conducted workshops including 80 attendants and a survey of 110 producers and users of systematic reviews. We calculated acceptability of statements and revised the wording of those that were unacceptable to ≥40% of participants.

**Results:** Most participants agreed statements should be based on size of effect and certainty of evidence. Statements for low, moderate and high certainty evidence were acceptable to >60%. Key guidance, for example, includes statements for high, moderate and low certainty for a large effect on intervention x as: x *results* in a large reduction…; x *likely results* in a large reduction…; x *may result* in a large reduction…, respectively.

**Conclusions:** Producers and users of systematic reviews found statements to communicate findings combining size and certainty of an effect acceptable. This article provides GRADE guidance and a wording template to formulate statements in systematic reviews and other decision tools.

**Keywords:** review literature as topic, health communication, Evidence-Based Medicine, Surveys and Questionnaires, Language, persuasive communication

**Running title:** Informative statements for systematic reviews

**Word count: 3582** (not including tables, boxes, figures or references)

**What is new?**

**Key findings**

- A set of statements to interpret results of systematic reviews of interventions and communicate them to patients, the public, and health care professionals was developed based on the GRADE approach to assess evidence. Experience with the statements and informal feedback showed that existing formulations were still not quite fit for purpose, and often used inconsistently.
- Building on results of workshops and a survey including producers and users of systematic reviews we revised the standardized statements.
- There was agreement that communicating the findings of reviews should be based on two components of a result: the magnitude or size of the effect and the certainty of the evidence.

**What this adds to what is known**

Inconsistent words and phrases have been used to communicate the results of systematic reviews to users. Our suggested standardized statements are informative and were found to be acceptable to producers and users of systematic reviews. We provide detailed guidance for how to use the statements.

**What is the implication, what should change now**

The template to formulate statements can be used to communicate the results of systematic reviews to users. These statements can be used in many sections of the systematic review, in evidence tables, and in tools or products for decision makers based on systematic reviews such as guideline recommendations.

## 1. Introduction

Systematic reviews aim to synthesise evidence and provide readers with a summary of the findings for a specific intervention. To achieve this goal, the findings should be communicated as clearly and as simply as possible. The GRADE approach posits that there are two important components of a result of a review: the effect of the intervention, presented as the risk or difference in effect, as absolute numbers (e.g., 5 fewer deaths per 100), or as a narrative synthesis; and the certainty of (or confidence in) the evidence for that effect (categorised using the GRADE approach into high, moderate, low and very low) [1-6]. Both components should be conveyed to avoid misleading the reader. Consider, for example, a systematic review of the effects of waiving surgical fees to improve the use of cataract surgical services [7]. The authors found a risk ratio of 1.94 for the uptake of surgery, which they determined was an important increase in uptake. The certainty of evidence was low due to indirectness and imprecision (95% CI 1.14 to 3.31). If the authors conclude that there is an increase in uptake, but do not indicate that there is low certainty, readers could misinterpret the result as meaning that waiving surgical fees does increase uptake when in fact there is uncertainty. Although, the levels of evidence provided by the GRADE approach should be used to communicate the results (e.g., there is moderate certainty evidence that intervention A has X effect), various other phrases have been used, such as 'limited evidence', 'insufficient evidence', 'no evidence to support', or 'the evidence shows, at best, a modest, non-statistically significant trend in favour of intervention A'. All of which can confuse readers. Previous research has explored methods to best communicate results and the GRADE Working Group has developed Evidence profiles and Summary of Findings Tables [3, 8-10]. While these tables help readers understand the results of

systematic reviews, this research found that many participants also appreciated brief statements describing the results [11, 12].

However, guidance for how to interpret and communicate results using statements is limited. The previous version of the Cochrane Handbook provided some guidance to not describe results as statistically or not statistically significant and avoid the common misinterpretation that large p values mean 'no difference' or 'no effect' or small p values mean an important effect [1, 8]. It also cautions authors about using 'evidence of no effect' or 'no evidence of effect' because these phrases are often used incorrectly. In 2010, we developed and tested four statements that were based on the size of an effect and the certainty of the evidence using the GRADE approach. Since then, we have received informal feedback suggesting that these statements are restrictive and other options are needed, and therefore we decided to improve and test new approaches.

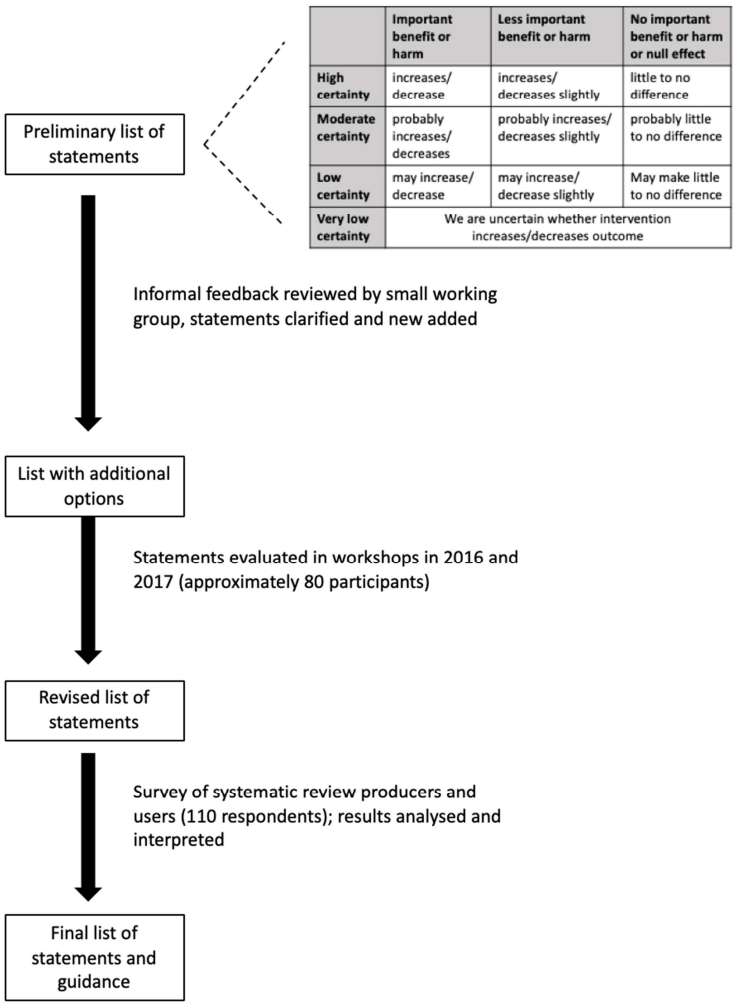
Our goal was to develop a set of standardized statements with multiple options for interpreting and communicating results of systematic reviews , and to write guidance. The statements assume that the evidence for an outcome is assessed using the GRADE approach or another formal system with four levels of evidence. It also assumes that certainty of evidence is not solely based on the imprecision of the result (i.e., power of the analysis and width of confidence interval), but also on other criteria, such as risk of bias of the studies, inconsistency (heterogeneity) of the result, indirectness (including subgroup analyses and applicability of the outcome measure), publication bias, and others.

## 2. Methods

### 2.1. Summary of research methods

The overall design is shown in Figure 1.

**Figure 1: Study design**




| | Important benefit or harm | Less important benefit or harm | No important benefit or harm or null effect |
|---|---|---|---|
| High certainty | increases/ decrease | increases/ decreases slightly | little to no difference |
| Moderate certainty | probably increases/ decreases | probably increases/ decreases slightly | probably little to no difference |
| Low certainty | may increase/ decrease | may increase/ decrease slightly | May make little to no difference |
| Very low certainty | We are uncertain whether intervention increases/decreases outcome | | |

### 2.2. Preliminary development

In 2010, during research to create a summary to present results from a systematic review to consumers, we developed, tested, and received feedback from an advisory group of statisticians about, statements to describe the effect of an intervention on an outcome. Single statements combined words for the size of an effect on an outcome and the certainty in that effect [12]. For example, suppose a review found that vitamin D results in an important reduction in falls with moderate certainty. The size of the effect would be described as *reduces*, and *probably* would indicate the certainty , and the final statement would be - "vitamin D probably reduces falls". Depending on the size/importance of the effect, different qualifiers were used: for an important reduction in an outcome, the verb used was *reduces*; to describe a less important effect *slightly reduces* was used; and when the effect was close to a null effect, *little to no difference* was used. A different qualifier was used to express certainty: high, moderate, low or very low certainty were conveyed as *will, probably, may,* and *we are uncertain*, respectively.

During this research, we explored different approaches. Initially, we had six different ways to categorise the size of an effect based on how wide/narrow the confidence intervals were. However, the width is already considered in the GRADE assessment and therefore the number of categories was reduced to three: *important, less important* and *little to no difference*. We also explored different qualifiers based on *why* evidence was rated down. If the evidence was low certainty because it was rated down twice for imprecision the qualifier was *we are very uncertain*, but if the evidence was rated down twice - once for imprecision and once for risk of bias - the qualifier was *possibly*. This system was after more discussion reduced to the four

categories of GRADE because the level of certainty reflects our uncertainty regardless of what specific domains are rated down.

### 2.3. Workshops

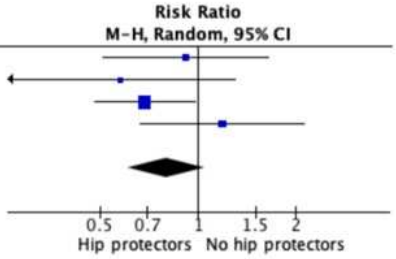
Following publication of the minimum set of statements and years of informal feedback, a small working group of authors met and created a longer list of options. We conducted three workshops at GRADE meetings in 2016 and 2017, each with approximately 20 – 40 people with expertise in methods of systematic reviews and guideline development, some of whom did not speak English as a first language. During the workshops, participants reviewed 4-6 examples of the results for an outcome of a systematic review as forest plot of a meta-analysis (Figure 2), a narrative synthesis, or in absolute effects, along with the certainty of the evidence and explanations We asked participants to discuss what statements they would use to express the result or if they agreed with the statement provided and why. We used the feedback to make revisions to our list.

**Figure 2: Example of information provided to workshop participants for feedback**
Note: the appropriate statement in this example is 'hip protectors probably reduces the risk of hip fractures slightly'

**Outcome: hip fractures**

| Study or Subgroup | Hip protectors Events | Total | No hip protectors Events | Total | Weight | Risk Ratio M-H, Random, 95% CI |
|---|---|---|---|---|---|---|
| Study 1 | 17 | 86 | 19 | 88 | 20.3% | 0.92 [0.51, 1.64] |
| Study 2 | 8 | 72 | 14 | 73 | 11.1% | 0.58 [0.26, 1.30] |
| Study 3 | 45 | 459 | 69 | 483 | 47.8% | 0.69 [0.48, 0.98] |
| Study 4 | 23 | 276 | 20 | 285 | 20.8% | 1.19 [0.67, 2.11] |
| **Total (95% CI)** | | **893** | | **929** | **100.0%** | **0.80 [0.61, 1.05]** |
| Total events | 93 | | 122 | | | |

Heterogeneity: $Tau^2 = 0.01$; $Chi^2 = 3.35$, df = 3 (P = 0.34); $I^2 = 10\%$
Test for overall effect: Z = 1.58 (P = 0.11)




A review of hip protectors versus no hip protectors to prevent hip fractures at 1 year found risk ratio (RR) 0.80 (95% CI, 0.61 to 1.05). This means in 100 people at high risk of hip fractures (10 per 100) that there would be 2 fewer fractures (from 4 fewer to 1 more). The evidence was assessed as moderate certainty because of lack of blinding in the studies and some imprecision due to few events. The authors indicated that any reduction in fractures is beneficial but 5 fewer is important.

The abstract included the statement "*hip protectors probably reduces the risk of hip fractures slightly*". Do you agree with this statement? Why or why not?

### 2.4. Survey

From March to April 2018, we conducted an electronic survey using SurveyMonkey to determine the acceptability of the statements (Appendix 1). We purposively invited by email: 1) people who conduct or summarise systematic reviews for use in decision making; 2) people who use systematic reviews; and 3) statisticians with systematic review experience. Members of the GRADE Working Group were also invited. Invited participants could forward the email to others and we sent one reminder one week later. The survey link was also sent via one author's professional Twitter account (approximately 2000 followers). The first part of the survey asked participants about their roles in reviews and epidemiological training. Section 2 presented results for one outcome from five systematic reviews with 3 to 4 statements. Respondents rated the statements as unacceptable, acceptable or ideal. Section 3 asked 'Do you agree in principle that conclusions should be based on the concepts of the importance/size of the effect and the certainty of the evidence?'. We piloted the survey in two people and revised

accordingly. The Hamilton Integrated Research Ethics Board waived formal ethics approval. One investigator analysed the data using descriptive statistics, and summarised the free-text comments by broad themes. A priori, we decided to revise statements that were 'unacceptable' to more than 40% and keep statements that more than 60% judged acceptable or ideal.

### 2.5. Incorporation of results

The lead authors incorporated the survey and workshop results into the statements and developed guidance. We presented the results to approximately 60 attendees at a GRADE Working Group meeting (April 2018) and to approximately 80 people in September 2018 (for approval.

## 3. Results and Implications

### 3.1 Acceptability of statements

Of the 110 respondents (19 of whom were members of this GRADE project group), 72% described themselves as systematic review or guideline methodologists, and 13% as readers of reviews. Approximately, 30% indicated they had no formal education in epidemiology. Two did not answer all questions; however, their results were included. In section 2, 39 provided written comments about acceptability, and 15 provided comments in section 3. We present results from the 91 participants and use the comments of the project members to contextualise results (see Appendix 2 for raw data from survey). We did not calculate a response rate since participants could forward the link to others. The final list of informative statements is in Table 1.

*Acceptability of statements for very low certainty evidence:* The statement "[Intervention X] may reduce the [outcome] slightly but we are uncertain" was presented in two examples and was rated as unacceptable by 37% in one example and 46% in the other. The comments highlighted that *we are uncertain* could be misinterpreted; respondents suggested that it would be clearer to instead write that *the evidence* is uncertain. The two examples also provided two statements stating the direction of effect: "We are uncertain about whether co-enzyme Q10 reduces blood pressure" – acceptable to 80%, and "We are uncertain about the effect of co-enzyme Q10 on blood pressure" – acceptable to 71%. During workshops, there was also some debate about communicating a direction of effect when the evidence is so uncertain. However, we have kept both options for very low certainty: uncertain effect with or without a direction of effect.

*Acceptability of statements for low certainty evidence:* Participants were presented with the qualifying words *may, appears, suggests, and likely* ("Probiotics may result in a large reduction in the incidence of diarrhea). *Likely* was rated as unacceptable by 52%; a*ppears* by 50%, and *suggests* by 57%. Respondents observed that most words to convey low certainty evidence were vague e.g., *may* could be interpreted may or may not. Respondents wrote that s*uggests* could be more acceptable, and some noted that *appears* sounded supernatural. Therefore, a*ppears* was deleted, but *may* and *suggests* remain options for low certainty evidence.

*Acceptability of statements for moderate certainty and high certainty evidence:* There were few comments and both *likely* and *probably* were acceptable.

*Acceptability of statements to communicate size of effect:* In one example, the intervention resulted in 2 more hip fractures per 1000 (from 2 fewer to 6 more) and the authors judged that 2 more did not reach a threshold for an effect either as a beneficial reduction or as a harm. Two of the example narrative statements used *results in little to no difference* and the other two used *does not reduce outcome*. *Little to no difference* was unacceptable to 20%, and *does not reduce* to 35-40%. There were many comments that *does not* should not be used when communicating a result close to null effect. Workshop participants also often expressed concern with interpreting null effect as *does not* affect.

Another example explored the acceptability of statements to convey evidence for a small effect that is not important. Two of the three statements describing the effect as a *small possible unimportant reduction* were rated as unacceptable by 45 to 50%. Participants responded that the high number of qualifying words could be confusing. Statements with multiple qualifiers for importance were therefore deleted and a *small effect* has been divided into a *small and important effect* and an unimportant effect as *trivial or small, unimportant or no effect ('trivial' is added to be consistent with GRADE's Evidence to Decision frameworks [13-17])*. In this example, *do not result in* was used and again there were comments that it is not correct to describe a result near the null effect as *not* occurring. The words *do not* or *does not* to describe *little to no effect* are still an option.

### 3.2 Agreement about principles of size of effect and certainty of evidence

Ninety-nine percent (84/85) agreed that statements should be based on both size of the effect and certainty of evidence. In general, respondents were concerned that it is difficult to determine whether an effect is large, moderate, small (important or not important), or of little to no effect. Comments also highlighted to not interpret wide confidence intervals and non-statistically significant results as *no effect*.

**Table 1: Final list of informative statements to communicate results of systematic reviews**

| **Size of the effect estimate** | **Suggested statements**<br>*(replace X with intervention, replace 'reduce/increase' with direction of effect, replace 'outcome' with name of outcome, include 'when compared with Y' when needed)* |
|---|---|
| **HIGH Certainty of the evidence** | |
| **Large effect** | X results in a large reduction/increase in outcome |
| **Moderate effect** | X reduces/increases outcome<br>X results in a reduction/increase in outcome |
| **Small important effect** | X reduces/increases outcome slightly<br>X results in a slight reduction/increase in outcome |
| **Trivial, small unimportant effect or no effect** | X results in little to no difference in outcome<br>X does not reduce/increase outcome |
| **MODERATE Certainty of the evidence** | |
| **Large effect** | X likely results in a large reduction/increase in outcome<br>X probably results in a large reduction/increase in outcome |
| **Moderate effect** | X likely reduces/increases outcome<br>X probably reduces/increases outcome<br>X likely results in a reduction/increase in outcome<br>X probably results in a reduction/increase in outcome |
| **Small important effect** | X probably reduces/increases outcome slightly<br>X likely reduces/increases outcome slightly<br>X probably results in a slight reduction/increase in outcome<br>X likely results in a slight reduction/increase in outcome |

| | |
|---|---|
| **Trivial, small unimportant effect or no effect** | X likely results in little to no difference in outcome<br>X probably results in little to no difference in outcome<br>X likely does not reduce/increase outcome<br>X probably does not reduce/increase outcome |
| **LOW Certainty of the evidence** | |
| **Large effect** | X may result in a large reduction/increase in outcome<br>The evidence suggests X results in a large reduction/increase in outcome |
| **Moderate effect** | X may reduce/increase outcome<br>The evidence suggests X reduces/increases outcome<br>X may result in a reduction/increase in outcome<br>The evidence suggests X results in a reduction/increase in outcome |
| **Small important effect** | X may reduce/increase outcome slightly<br>The evidence suggests X reduces/increases outcome slightly<br>X may result in a slight reduction/increase in outcome<br>The evidence suggests X results in a slight reduction/increase in outcome |
| **Trivial, small unimportant effect or no effect** | X may result in little to no difference in outcome<br>The evidence suggests that X results in little to no difference in outcome<br>X may not reduce/increase outcome<br>The evidence suggests that X does not reduce/increase outcome |
| **VERY LOW Certainty of the evidence** | |
| **Any effect** | The evidence is very uncertain about the effect of X on outcome<br>X may reduce/increase/have little to no effect on outcome but the evidence is very uncertain |

## 4. Discussion and Guidance

### 4.1. Discussion

We have created a list of brief and informative statements that authors of systematic reviews, and people presenting evidence to decision makers, e.g., guideline developers, can use to describe the results (Table 1). This work builds on our previous research, on many years of experience using the statements, a survey, and on feedback received during GRADE working group meetings. Although we piloted examples and the survey, there is still the potential that we may not have expressed the task clearly to respondents, resulting in some confusion. However, we received comments from a variety of important stakeholders, including

methodologists in systematic reviews and guidelines and readers, and found results were consistent. We provide guidance to use these statements, and examples in Appendix 3.

### 4.2. Use of certainty of evidence and size of effect to write informative statements

The basic premise is that review authors should report both the effect of an intervention on an outcome and the certainty in the evidence. Authors can communicate these components in multiple ways. GRADE guidance now suggests two approaches. First, authors may communicate the findings by providing the effect on the outcome and the certainty of the evidence according to the GRADE levels of evidence (i.e., provide the point estimate and confidence interval in relative and absolute terms, and then specify that the evidence is "moderate certainty"). Second, if authors want to communicate the result in one statement, they should use Table 1, first selecting the category for certainty of evidence, then making a judgement regarding the size of the effect, and finally choosing from the appropriate wording options (e.g., for a small important effect of moderate certainty - "intervention A likely increases outcome X slightly").”

### 4.3. Decisions about the size of the effect

To create a statement using Table 1, authors must decide into which category the size of effect falls. The GRADE Evidence to Decision framework provides some guidance about the size of effect [13-17]. However, when conducting a GRADE assessment, in particular when assessing imprecision, systematic reviewers partially contextualise decisions using thresholds for *no or trivial, small, moderate and large effects* [18-21]. These decisions can be based on research into minimal important differences, discussions within the systematic review team, or consultation

with decision-makers, and should be transparent. Two considerations are of critical importance when determining the size. The first is calculating and using absolute effects rather than using relative effects that can often be misleading. For instance, consider a risk ratio 0.84, or 16% relative reduction in hip fractures in older adults. If on the one hand, the baseline risk of hip fractures is 20/1000 over 1 year, the risk ratio 0.84 would translate to 3 fewer per 1000, which most would consider a small effect. On the other hand, if the baseline risk is 200/1000, many would consider that the resulting absolute reduction of 32 per 1000 is a moderate to large effect. The second is identifying the value of the outcome [16, 17]. Ideally, review authors identify the thresholds, and use them to rate the certainty of the evidence. The approach to choose a threshold (or range) can be either fully contextualised (based on consideration of all critical outcomes) or partially contextualised (based on the value of the individual outcome.) [20]. Whatever the thresholds, a decision needs to be made in order to write a statement using Table 1.

When deciding on thresholds, review authors also need to be aware of the risk of misinterpreting a result with a wide confidence interval that includes '1' (for relative effects) or '0' (for absolute effects) as 'no effect' or 'no difference' [22, 23]. For example, consider a mean difference for the effect of a treatment on quality of life is 1.5 (95% CI, -1.2 to 4.2) where an important effect is an increase of 1 on a scale of 1 to 10 (better), and the certainty of the evidence is low (due to imprecision and risk of bias). The point estimate is an increase of 1.5, and we would characterise the effect as important, likely moderate, but not 'no effect'. Authors need to determine the size of the effect based on the effect estimate, not on the confidence

intervals. The width of the confidence interval is considered in the assessment of the certainty of the evidence (see Box 1). In this case, the certainty is low,we use the word 'may', and the final statement is, 'the [treatment] may increase quality of life'. In contrast, if the effect was an increase of 0.3 (95% CI, -1.8 to 2.3), the effect could be categorised as 'trivial, small unimportant or no effect' because the effect estimate is less than our threshold for an important difference, and the final statement based on low certainty evidence would be 'the [treatment] may have little to no effect on quality of life'.

Box 1: Best estimate versus confidence intervals to determine effect size

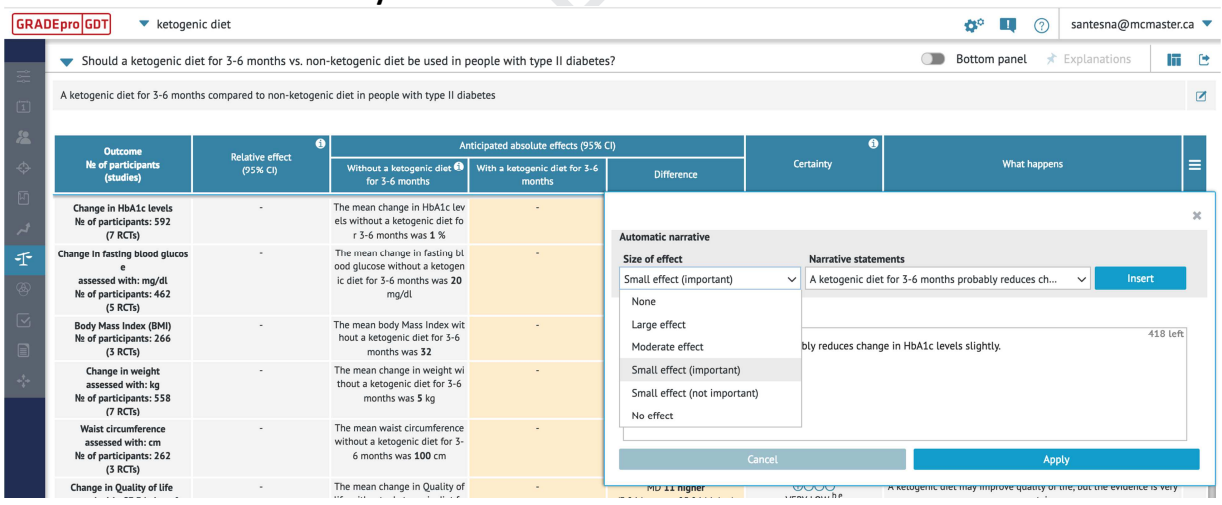
The statements communicate the size of the effect based on the point estimate in a meta-analysis or on the summary estimate in a narrative synthesis instead of the confidence intervals. Confidence intervals represent the range in which a point estimate would fall if multiple experiments were conducted, or as the range of values either side of the estimate between which we can be 95% sure that the true value lies [22], and are calculated based on factors such as sample sizes and variance within or between studies. The calculation does not factor in the risk of bias of the studies; indirectness of the populations, interventions or outcomes; or, the risk of publication bias, which (if there were methods to do so) could widen the confidence intervals, making the calculated confidence intervals meaningless. However, when conducting a GRADE assessment authors consider the width of the confidence intervals and power of the analysis (i.e., imprecision) plus all of the other factors to determine the certainty of the evidence. Thus, the certainty around the point estimate varies depending on what domains demonstrate shortcomings and except for imprecision that certainty interval is not known [18, 19]. For this reason, when communicating an effect using statements, authors

should focus on the best estimate and on the certainty in that estimate which considers multiple factors.

### 4.4. Use the statements in the text of a review and in Summary of Findings tables

Authors can use these statements throughout a systematic review: in the abstract, plain language summary, results, discussion, and in evidence tables. Experience has shown that this approach to wording should not be an automated application, which could result in a list of monotonous statements. In GRADEpro (www.gradepro.org), the software programme to produce summary of findings tables, the size of effect and the certainty of evidence are used to automatically generate an editable statement (Figure 3).

**Figure 3: Screenshot of GRADEpro and automatic generation of informative statements based on size of effect and certainty of evidence**




Systematic reviews typically compare an intervention/test to a comparator. The statements in Table 1 do not explicitly state the comparator which may be acceptable when the comparator is standard care, a placebo, or no intervention, but when it is an alternative intervention, it's important to include it. Using a hypothetical example, there is low certainty evidence that oseltamivir reduced the duration of symptoms by 2 days (95% CI, 0.5 to 3.6 days) when

compared to zanamivir, whereby 2 days was an important difference. The informative statement should be 'oseltamivir may reduce the duration of symptoms more than zanamivir'.

**4.5 Borderline decisions and very low certainty of the evidence**

When applying the GRADE approach, authors may debate about the weight of each domain to determine the level of evidence. For example, in some cases, moderate certainty evidence may be due solely to imprecision, in other cases, it may be a combination of small concerns with imprecision, risk of bias and inconsistency. Despite these differences, authors must make a final decision about the level of evidence, and it is this level that determines the wording options available to use in that category. The GRADE approach to certainty of evidence, however, acknowledges that, despite the four categories of high, moderate, low and very low, certainty is a continuum [2]. Consequently, users may find that when deciding on the certainty they may have been on the threshold between categories, but ultimately had to choose a category, make a borderline decision, or characterise the certainty as being at a threshold. When choosing a statement in these instances, users could choose from the statements on either side of the border.

We have also provided two options for a statement based on very low certainty of evidence: one option gives the direction of the effect, the other does not. Ratings are on a continuum and within the category of very low there may be situations when authors feel somewhat more compelled to express an effect (e.g., when the rating borders on low) and situations when they do not (e.g. the evidence is at the very bottom of the continuum of certainty).

### 4.6. Use of the statements in different review types

The underlying principle considering size of effect and certainty of evidence (whether GRADE or another system with four levels) to write statements can likely be applied to any review type. In a test accuracy review with pooled sensitivity and specificity estimates, the absolute numbers of misidentified people (i.e., false negatives and positives) can be quantified as large, moderate, small, or trivial, depending on the consequences for patients. A review may find that a cytology test misses 20 more out of 1000 women with cervical cancer lesions than an HPV test - a small difference based on moderate certainty evidence. We could conclude that 'when compared to HPV tests, cytology tests *probably* miss *slightly* more women with cervical lesions.' In prognostic reviews, the statements could be written as 'associations'. For example, for a moderately sized association of hip fractures with age and low certainty evidence, the statement would be 'age *may* be associated with hip fractures'.

## 5. Conclusions

The informative statements to communicate results of systematic reviews should be used throughout the text of a systematic review, in the abstract, plain language summary, results, discussion, and in evidence tables. These statements can also be used in other tools and products that communicate the results of systematic reviews to decision makers, and in fact are already being used in health care guidelines to summarise the evidence and in patient versions of guidelines [27-29]. The list was also originally translated into Spanish, Norwegian, Italian, French and German [12], and future work will focus on these translations.

**Acknowledgements**

We would also like to acknowledge specific GRADE Working Group members that provided help with the project: Arnav Agarwal, Sarah Rosenbaum, Jasvinder Singh, Airton Stein, Judith Thornton, Gemma Villanueva, and Lee Yee Chong.

**References**


1. Schünemann HJ OA, Vist GE, Higgins JPT, Deeks JJ, Glasziou P, Guyatt GH: **Chapter 12: Interpreting results and drawing conclusions.** In: *Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011).* Edited by Higgins JPT GS. Available from www.cochrane-handbook.org: The Cochrane Collaboration, 2008; 2008.
2. Guyatt G, Oxman AD, Sultan S, Brozek J, Glasziou P, Alonso-Coello P, Atkins D, Kunz R, Montori V, Jaeschke R *et al*: **GRADE guidelines 11-making an overall rating of confidence in effect estimates for a single outcome and for all outcomes**. *J Clin Epidemiol* 2012.
3. 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**. *J Clin Epidemiol* 2011, **64**(4):383-394.
4. Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knotterus A: **GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology**. *J Clin Epidemiol* 2010.
5. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, Schunemann HJ: **GRADE: an emerging consensus on rating quality of evidence and strength of recommendations**. *BMJ* 2008, **336**(7650):924-926.
6. Schunemann HJ, Best D, Vist G, Oxman AD: **Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations**. *CMAJ* 2003, **169**(7):677-680.
7. Ramke J, Petkovic J, Welch V, Blignault I, Gilbert C, Blanchet K, Christensen R, Zwi AB, Tugwell P: **Interventions to improve access to cataract surgical services and their impact on equity in low- and middle-income countries**. *The Cochrane database of systematic reviews* 2017, **11**:Cd011307.
8. Schünemann HJ HJ, Vist GE, Glasziou P, Akl E, Skoetz N, Guyatt GH: **Chapter 14: Completing Summary of findings tables and grading the certainty of evidence**. In: *Cochrane Handbook for Systematic Reviews of Interventions Version 6 (updated January 29, 2019).* Edited by Higgins JPT TJ, Chandler J, Cumston M, Li T, PageMJ, Welch V. Available from https://training.cochrane.org/handbooks: The Cochrane Collaboration, 2019; 2019.
9. Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, Brozek J, Norris S, Meerpohl J, Djulbecovic B *et al*: **GRADE guidelines 12. Preparing Summary of Findings tables-binary outcomes**. *J Clin Epidemiol* 2012.
10. Schünemann HJ OA, Higgins JPT, Vist GE, Glasziou P, Guyatt GH: **Chapter 11: Presenting results and 'Summary of findings' tables**. In: *Cochrane Handbook for Systematic Reviews of Interventions Version 510 (updated March 2011).* Edited by Higgins JPT GS. Available from www.cochrane-handbook.org: The Cochrane Collaboration, 2008; 2008.
11. Carrasco-Labra A, Brignardello-Petersen R, Santesso N, Neumann I, Mustafa RA, Mbuagbaw L, Etxeandia Ikobaltzeta I, De Stio C, McCullagh LJ, Alonso-Coello P *et al*: **Improving GRADE evidence tables part 1: a randomized trial shows improved understanding of content in summary of findings tables with a new format**. *J Clin Epidemiol* 2016, **74**:7-18.

12. Glenton C, Santesso N, Rosenbaum S, Nilsen ES, Rader T, Ciapponi A, Dilkes H: **Presenting the results of Cochrane Systematic Reviews to a consumer audience: a qualitative study**. *Med Decis Making* 2010, **30**(5):566-577.
13. Moberg J, Oxman AD, Rosenbaum S, Schunemann HJ, Guyatt G, Flottorp S, Glenton C, Lewin S, Morelli A, Rada G *et al*: **The GRADE Evidence to Decision (EtD) framework for health system and public health decisions**. *Health Res Policy Syst* 2018, **16**(1):45.
14. Schunemann HJ, Wiercioch W, Brozek J, Etxeandia-Ikobaltzeta I, Mustafa RA, Manja V, Brignardello-Petersen R, Neumann I, Falavigna M, Alhazzani W *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**:101-110.
15. Parmelli E, Amato L, Oxman AD, Alonso-Coello P, Brunetti M, Moberg J, Nonino F, Pregno S, Saitto C, Schunemann HJ *et al*: **GRADE EVIDENCE TO DECISION (EtD) FRAMEWORK FOR COVERAGE DECISIONS**. *Int J Technol Assess Health Care* 2017:1-7.
16. Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, Treweek S, Mustafa RA, Rada G, Rosenbaum S *et al*: **GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction**. *BMJ* 2016, **353**:i2016.
17. Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, Treweek S, Mustafa RA, Vandvik PO, Meerpohl J *et al*: **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.
18. Schunemann HJ: **Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision?** *J Clin Epidemiol* 2016, **75**:6-15.
19. Anttila S, Persson J, Vareman N, Sahlin NE: **Conclusiveness resolves the conflict between quality of evidence and imprecision in GRADE**. *J Clin Epidemiol* 2016, **75**:1-5.
20. Hultcrantz M, Rind D, Akl EA, Treweek S, Mustafa RA, Iorio A, Alper BS, Meerpohl JJ, Murad MH, Ansari MT *et al*: **The GRADE Working Group clarifies the construct of certainty of evidence**. *Journal of clinical epidemiology* 2017, **87**:4-13.
21. 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**. *Journal of clinical epidemiology* 2011, **64**(12):1283-1293.
22. Altman DG: **Why we need confidence intervals**. *World journal of surgery* 2005, **29**(5):554-556.
23. Nuzzo R: **Scientific method: statistical errors**. *Nature* 2014, **506**(7487):150-152.
24. Zhang Y, Coello PA, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, Pardo-Hernandez H, Etxeandia-Ikobaltzeta I, Qaseem A, Williams JW, Jr. *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* 2018.
25. Zhang Y, Alonso-Coello P, Guyatt GH, Yepes-Nunez JJ, Akl EA, Hazlewood G, Pardo-Hernandez H, Etxeandia-Ikobaltzeta I, Qaseem A, Williams JW, Jr. *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* 2018.

26. Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, McGinn T, Hayden J, Williams K, Shea B *et al*: **Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients**. *BMJ* 2015, **350**:h870.
27. Papaioannou A, Santesso N, Morin SN, Feldman S, Adachi JD, Crilly R, Giangregorio LM, Jaglal S, Josse RG, Kaasalainen S *et al*: **Recommendations for preventing fracture in long-term care**. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2015, **187**(15):1135-1144, e1450-1161.
28. Santesso N, Carrasco-Labra A, Brignardello-Petersen R: **Hip protectors for preventing hip fractures in older people**. *The Cochrane database of systematic reviews* 2014(3):Cd001255.
29. Wieland LS, Santesso N: **A Summary of a Cochrane Review: Acupuncture or acupressure for induction of labour**. *European journal of integrative medicine* 2018, **17**:141-142.

## APPENDIX 1: SURVEY

**Thank you for volunteering to participate in our survey.**

**This survey is being conducted by a project group of the GRADE Working Group. The Hamilton Integrated Research Ethics Board (HiREB) at McMaster University has waived the requirement for individual consent. Your answers to this survey will be entered into a large database and will remain confidential and anonymous (unless you provide your name so that we can contact you for verbal feedback). This survey should take about 15-20 minutes to complete.**

**We have been working on single statements to communicate the results of a systematic review. Some example statements could be "taking vitamin C daily probably reduces your risk of catching a cold" or "exercising 2 hours a week increases sleep duration slightly". How we write these statements are based on the importance/size of the effect (e.g., minimally important difference, thresholds), and the certainty of the evidence.**

**In this survey, you will be shown 5 examples of the results of a systematic review and asked how acceptable you think the single statement is. You can complete all 5 examples or stop any time. You will also be given space to provide general comments about the statements at the end.**

**The survey is not a test at all. Rather, we would really like you to provide your opinion about acceptable ways to communicate the results.**

**If you have any questions about this survey, please contact Nancy Santesso at santesna@mcmaster.ca**

## Background Information

* 1. What is your primary role related to systematic reviews and guidelines? (Select one that best represents your role)

- ◯ Methodologist who conducts systematic reviews
- ◯ Clinical expert who conducts systematic reviews
- ◯ Methodologist who has been involved in guideline development
- ◯ Clinical expert who has been involved in guideline development
- ◯ Methodologist not involved in systematic reviews or guidelines
- ◯ Clinician not involved in systematic reviews or guidelines
- ◯ Someone who reads systematic reviews

2. What is your education in epidemiology?

3. A systematic review compared the effects of cognitive behavioural therapy versus a waiting list for military suffering from post-traumatic stress disorder on depression.
It found that cognitive behaviour therapy reduced depression by 8 points more on a scale from 1-100 (95% confidence interval from 21 point reduction to 12 point increase). This reduction is small but important.
The evidence came from a meta-analysis with very few people (91) and very serious concern that the studies were at high risk of bias because of unclear randomisation and large loss to follow-up.

The conclusion about the effect of cognitive behaviour therapy could be worded in the following three ways. Please indicate the acceptability of each statement.

| | Unacceptable | Acceptable | Ideal |
|---|---|---|---|
| a. Cognitive behaviour therapy may reduce depression slightly more than no therapy but we are uncertain. | ◯ | ◯ | ◯ |
| b. We are uncertain about the effect of cognitive behaviour therapy on depression. | ◯ | ◯ | ◯ |
| c. We are uncertain about whether cognitive behaviour therapy reduces depression more than no therapy. | ◯ | ◯ | ◯ |

A systematic review compared the effects of co-enzyme Q10 versus placebo on blood pressure. A summary of the findings is provided below. Note that the 1.62 mm/Hg reduction in blood pressure is a small but important effect, but the evidence was assessed at 'very low quality/certainty'.

| What was measured | Without co-enzyme Q10 | With co-enzyme Q10 | Quality of the evidence[b] |
|---|---|---|---|
| Diastolic blood pressure (2 studies, 71 people) | 3 mm/Hg lower | Lower by 1.62 mm/Hg more (from 5.20 lower to 1.96 higher)[a] | ⊕⊖⊖⊖ Very low |

[b] Details about the quality of the evidence: evidence was very low quality because it is unclear if the studies were well-conducted and there were very few people in the studies.

4. The conclusion about the effect of co-enzyme Q10 on blood pressure could be worded in the following three ways. Please indicate the acceptability of each statement.

| | Unacceptable | Acceptable | Ideal |
|---|---|---|---|
| a. Co-enzyme Q10 may reduce blood pressure slightly but we are uncertain. | ○ | ○ | ○ |
| b. We are uncertain about the effect of co-enzyme Q10 on blood pressure | ○ | ○ | ○ |
| c. We are uncertain about whether co-enzyme Q10 reduces blood pressure. | ○ | ○ | ○ |

Please see the results of a systematic review of probiotics compared to placebo on the incidence of diarrhea in children.

**Probiotics as an adjunct to antibiotics for the prevention of pediatric antibiotic-associated diarrhea in children**

**Patient or population:** children given antibiotics
**Settings:** inpatients and outpatients
**Intervention:** probiotics

| Outcomes | Illustrative comparative risks* (95% CI) | | Relative effect (95% CI) | No of Participants (studies) | Quality of the evidence (GRADE) |
|---|---|---|---|---|---|
| | Assumed risk | Corresponding risk | | | |
| | Control | Probiotics | | | |
| **Incidence of Diarrhea: Probiotic dose (equal to/ greater than) 5 billion CFU/day** Follow-up: 10 days to 3 mo's | **223 per 1000** | **89 per 1000** (65 to 122) | **RR 0.4** (0.29 to 0.55) | 1474 (7 studies) | ⊕⊕◯◯ **low**[1,2] |

[1] 2 of 7 trials had a high risk of bias due to high loss to follow-up (29% for both). Furthermore, loss to follow-up across the 7 trials was also high (16%)
[2] Sparse data (225 events) and the 95% CI for the extreme-plausible ITT analysis (60% of children loss to follow-up in probiotic group and 20% loss to follow-up in the control group had diarrhea) is wide (touches the line of 1) which also indicates imprecision (RR 0.72; 95% CI 0.53 to 0.99; $I^2 = 57\%$; $P = .04$)

5. The authors of the review considered that the cut-off for a large effect is RR 0.60. Please indicate the acceptability of the statements below.

| | Unacceptable | Acceptable | Ideal |
|---|---|---|---|
| a. Probiotics may result in a large reduction in the incidence of diarrhea. | ◯ | ◯ | ◯ |
| b. Probiotics likely result in a large reduction in the incidence of diarrhea. | ◯ | ◯ | ◯ |
| c. Probiotics appear to result in a large reduction in the incidence of diarrhea. | ◯ | ◯ | ◯ |
| d. The evidence suggests that probiotics result in a large reduction in the incidence of diarrhea. | ◯ | ◯ | ◯ |

Please see the results of a systematic review on the number of hip fractures older people living in the community experience when wearing hip protectors or not wearing hip protectors. Hip protectors are cushioned undergarments that could deflect or cushion the impact of a fall.

**Provision of hip protectors for preventing hip fractures in older people**

**Patient or population:** older people
**Settings:** institutional and community settings
**Intervention:** provision of hip protectors

| Outcomes | Illustrative comparative risks* (95% CI) | | Relative effect (95% CI) | No of Participants (studies) | Quality of the evidence (GRADE) |
|---|---|---|---|---|---|
| | No hip protectors provided | Hip protectors provided | | | |
| **Hip fractures at 1 year** older people living in the community Follow-up: 6-28 months | **Moderate risk**[1] **10 per 1000** | **12 per 1000** (8 to 16) | **RR 1.15** (0.84 to 1.58) | 5614 (5 studies) | ⊕⊕⊕○ **moderate**[2] |

[1] Median risk in people not provided with hip protectors across randomised controlled trials.
[2] Participants were not blinded and results are imprecise due to few reported events; however, baseline risk and absolute effects are small, therefore quality of the evidence was only downgraded from high to moderate quality.

6. The authors indicate that the effect found was less than their cut-off for an effect.

Please indicate the acceptability of the statements to communicate the effects of hip protectors compared to no hip protectors on the number of hip fractures.

| | Unacceptable | Acceptable | Ideal |
|---|---|---|---|
| a. Hip protectors likely do not reduce hip fractures. | ○ | ○ | ○ |
| b. Hip protectors likely result in little to no difference in hip fractures. | ○ | ○ | ○ |
| c. Hip protectors probably do not reduce hip fractures. | ○ | ○ | ○ |
| d. Hip protectors probably result in little to no difference in hip fractures. | ○ | ○ | ○ |

Please see below the results for the effect of oral leukotriene receptor antagonists on daytime nasal symptoms compared to placebo.

**Question:** Should oral leukotriene receptor antagonists vs placebo be used for treatment of seasonal allergic rhinitis?
**Bibliography:** Rodrigo G.J., Yanez A. The role of antileukotriene therapy in seasonal allergic rhinitis: a systematic review of randomized trials. Ann Allergy Asthma Immunol, 2006;96:779-786.

| Quality assessment | | | | | | | Summary of findings | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | No of patients | | Effect | | Quality |
| No of studies | Design | Limitations | Inconsistency | Indirectness | Imprecision | Other considerations | oral leukotriene receptor antagonists | placebo | Relative (95% CI) | Absolute | |
| Daytime nasal symptoms (follow-up 2 to 4 weeks; Better indicated by less) | | | | | | | | | | | |
| 6 | randomised trial | no serious limitations | no serious inconsistency | no serious indirectness | no serious imprecision | none | 2512[1] | 2512[1] | - | SMD -0.24 (-0.16 to -0.33) | ⊕⊕⊕⊕ HIGH |

7. Please indicate the acceptability of the statements below to communicate the effect of antagonists compared to placebo on daytime nasal symptoms.

*Note: the authors considered the SMD and confidence interval a small but not important effect.*

| | Unacceptable | Acceptable | Ideal |
|---|---|---|---|
| a. Antagonists result in a small effect that may not be an important reduction in daytime symptoms. | ○ | ○ | ○ |
| b. Antagonists result in a small possible unimportant reduction in daytime symptoms. | ○ | ○ | ○ |
| c. Antagonists do not result in an important reduction in daytime symptoms. | ○ | ○ | ○ |

8. Please take 5 minutes to review this list of options for statements to communicate results at various levels of evidence and size of effect. If you'd like, please provide any general comments about the statements below.

**Standardised wording of results and interpretation**

Note: the outcome mortality is used as an example and a reduction is desired; substitute 'mortality' for your outcome name and the direction of desired effect (e.g., reduce, increase, improve)

| Size of effect | Suggested wording |
|---|---|
| **Certainty of evidence: HIGH (does, results in, will)** | |
| Large effect (it's always based on importance of the outcome and size of the effect) | X results in a large reduction in mortality |
| Moderate effect | X reduces mortality<br>X results in a reduction in mortality |
| Small effect (important) | X reduces mortality slightly |
| Small effect (not important) | X results in a small effect that may not be an important reduction in mortality<br>X results in a small possibly unimportant effect in mortality<br>X does not result in an important reduction in mortality |
| No effect | X does not reduce mortality<br>X results in little to no difference in mortality |
| **Certainty of evidence: MODERATE (likely, probably)** | |
| Large effect | X likely results in a large reduction in mortality<br>X probably results in a large reduction in mortality |
| Moderate effect | X likely reduces mortality<br>X probably reduces mortality |
| Small effect (important) | X probably reduces mortality slightly |
| Small effect (not important) | X likely results in a small effect that may not be an important (or unimportant) reduction in mortality<br>X likely results in a small possibly unimportant effect in mortality<br>X probably results in a small effect that may not be an important (or unimportant) reduction in mortality<br>X probably results in a small possibly unimportant effect in mortality |
| No effect | X likely does not reduce mortality<br>X likely results in little to no difference in mortality<br>X probably does not reduce mortality<br>X probably results in little to no difference in mortality |
| **Certainty of evidence: LOW (may, suggests, appears to)** | |
| Large effect | X may result in a large reduction in mortality<br>X appears to result in a large reduction in mortality<br>The evidence suggests that X results in a large reduction in mortality |
| Moderate effect | X may reduce mortality<br>X appears to reduce mortality<br>The evidence suggests X reduces mortality |
| Small effect (important) | X may reduce mortality slightly<br>X appears to reduce mortality slightly<br>The evidence suggests X reduces mortality slightly |
| Small effect (not important) | X may result in a small effect that may not be an important (or unimportant) reduction in mortality<br>X appears to result in a small effect that may not be an important (or unimportant) reduction in mortality<br>The evidence suggests that X results in a small effect that may not be an important (or unimportant) reduction in mortality<br>X may result in a small possibly unimportant effect in mortality<br>X may result in a small effect that may not be an important (or unimportant) reduction in mortality<br>X may result in a small possibly unimportant effect in mortality |
| No effect | X may not reduce mortality<br>X may result in little to no difference in mortality<br>X appears to not reduce mortality<br>X appears to result in little to no difference in mortality<br>The evidence suggests that X does not reduce mortality |
| **Certainty of evidence: VERY LOW** | |
| Effect | X may reduce mortality but we are very uncertain<br>We are uncertain about the effect of X on mortality<br>We are uncertain about whether X reduces mortality |

9. If you would like to provide additional comments verbally, please provide your contact information and we will contact you.

10. One last question:

Do you agree in principle that conclusions should be based on the concepts of the importance/size of the effect and the certainty of the evidence?

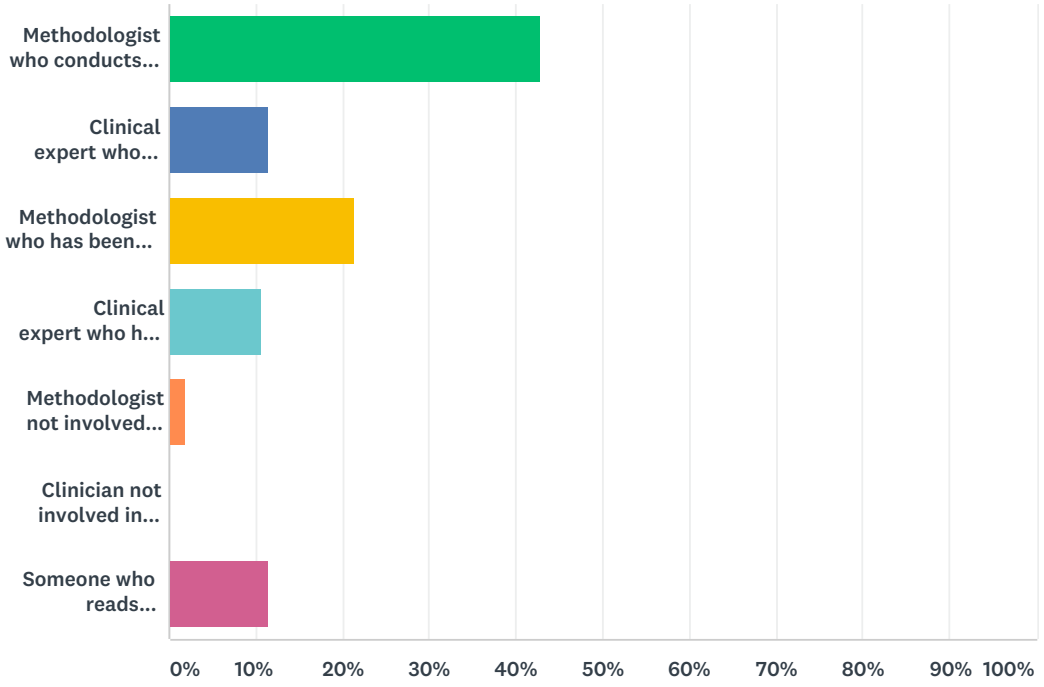
- ◯ Yes
- ◯ No

Please provide comments if you'd like.

**APPENDIX 2: Results of the survey of participants (not including GRADE project group members)**

## Q1 What is your primary role related to systematic reviews and guidelines? (Select one that best represents your role)

Answered: 112 Skipped: 0




| ANSWER CHOICES | RESPONSES | |
|---|---|---|
| Methodologist who conducts systematic reviews | 42.86% | 48 |
| Clinical expert who conducts systematic reviews | 11.61% | 13 |
| Methodologist who has been involved in guideline development | 21.43% | 24 |
| Clinical expert who has been involved in guideline development | 10.71% | 12 |
| Methodologist not involved in systematic reviews or guidelines | 1.79% | 2 |
| Clinician not involved in systematic reviews or guidelines | 0.00% | 0 |
| Someone who reads systematic reviews | 11.61% | 13 |
| TOTAL | | 112 |

# Q2 What is your education in epidemiology?

Answered: 87 Skipped: 25

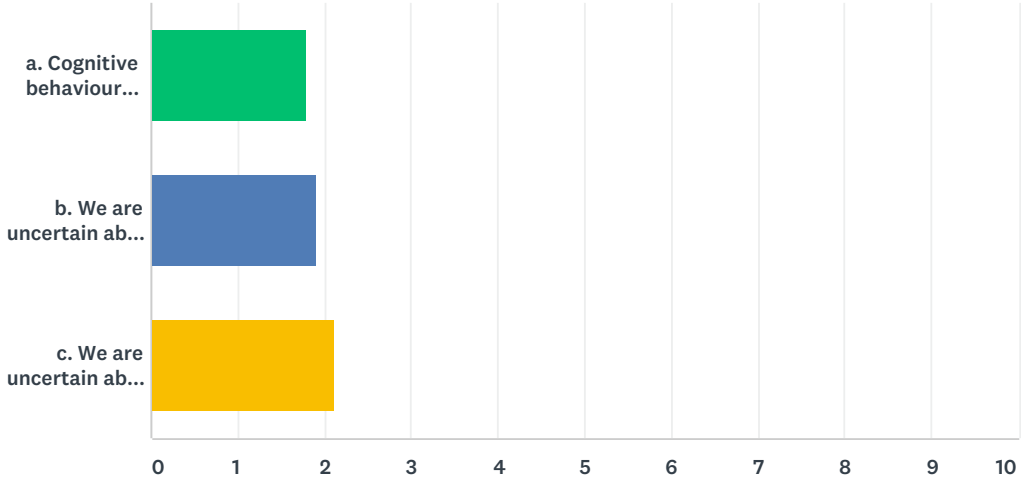
| # | RESPONSES | DATE |
|---|---|---|
| 1 | MSc epidemiology | 4/12/2018 12:45 PM |
| 2 | Masters | 4/9/2018 5:02 PM |
| 3 | Minimal; a class or two. | 4/9/2018 9:39 AM |
| 4 | little, basic courses | 4/9/2018 4:36 AM |
| 5 | MSc Doctor of Medicine [UK] | 4/8/2018 8:02 PM |
| 6 | undergraduate level formal teaching as well as experience and knowledge gained in the workplace. | 4/8/2018 3:13 AM |
| 7 | MSc | 4/6/2018 12:10 PM |
| 8 | no formal education | 4/6/2018 5:11 AM |
| 9 | Masters | 4/5/2018 3:26 PM |
| 10 | PhD | 4/5/2018 12:19 PM |
| 11 | MSc Epidemiology | 4/4/2018 8:08 AM |
| 12 | One post-graduate level paper | 4/4/2018 4:43 AM |
| 13 | Post graduate master epidemiology. | 4/4/2018 4:41 AM |
| 14 | PhD in statistical modelling | 4/4/2018 3:31 AM |
| 15 | One year training in clinical epidemiology. Autodidact learning and teaching. Participating in guideline development and implementation and in knowledge transfer. | 4/3/2018 9:51 PM |
| 16 | phd | 4/3/2018 9:11 AM |
| 17 | Master in Public Health | 4/3/2018 8:40 AM |
| 18 | One semester Phd course. | 4/3/2018 6:35 AM |
| 19 | None | 4/3/2018 3:26 AM |
| 20 | As part of MPH degree | 4/3/2018 3:25 AM |
| 21 | PhD | 4/2/2018 11:28 PM |
| 22 | PhD | 4/2/2018 2:50 PM |
| 23 | PhD epidemiology and biostats | 4/2/2018 2:11 PM |
| 24 | MSc in Clinical Epidemiology | 4/2/2018 7:35 AM |
| 25 | MSc | 4/2/2018 2:41 AM |
| 26 | none | 4/1/2018 7:14 AM |
| 27 | Diploma in evidence based healthcare | 3/31/2018 10:30 PM |
| 28 | basic | 3/31/2018 3:28 PM |
| 29 | MSc Clinical Epidemiology and Biostatistics | 3/30/2018 5:34 PM |
| 30 | MSc Clinical Epidemiology | 3/30/2018 11:55 AM |
| 31 | Masters in Public Health | 3/29/2018 6:32 PM |
| 32 | Master of Public Health from Columbia University | 3/29/2018 4:47 PM |
| 33 | Not sure how to answer this. Do you mean how many courses we had as master's and doctoral students we have had? I have had training in classical epidemiology and clinical epidemiology. | 3/29/2018 4:05 PM |

| | | |
|---|---|---|
| 34 | None. Masters in psychology. | 3/29/2018 2:34 PM |
| 35 | 15 hrs online course | 3/29/2018 1:42 PM |
| 36 | No formal | 3/29/2018 12:35 PM |
| 37 | PhD in nursing | 3/29/2018 12:15 PM |
| 38 | Master in Public Health | 3/29/2018 9:18 AM |
| 39 | Several courses as part of education. | 3/29/2018 8:22 AM |
| 40 | Phd | 3/29/2018 8:08 AM |
| 41 | I have a PhD in epidemiology and teach epidemiology for medical students. | 3/29/2018 7:34 AM |
| 42 | PhD | 3/29/2018 6:19 AM |
| 43 | during my PhD I became MSc in the epidemiology within a two year programma | 3/29/2018 6:01 AM |
| 44 | Undergraduate level. I now work in communication of data. | 3/29/2018 5:40 AM |
| 45 | None. | 3/29/2018 2:04 AM |
| 46 | PhD, with over 10 years experience following | 3/28/2018 9:59 PM |
| 47 | None formally. 20 years of experience with Cochrane | 3/28/2018 7:03 PM |
| 48 | Master in clinical and translational science | 3/28/2018 6:51 PM |
| 49 | MSc | 3/28/2018 4:31 PM |
| 50 | msc | 3/28/2018 4:12 PM |
| 51 | Training in Clinical epidemiology 450 h | 3/28/2018 4:05 PM |
| 52 | Masters in Health Science specialising in critical appraisal of clinical practice guidelines | 3/28/2018 1:29 PM |
| 53 | MSc Health Science. Post-grad training in population health. | 3/28/2018 1:22 PM |
| 54 | PhD in public health | 3/28/2018 11:44 AM |
| 55 | Master in Public Health | 3/28/2018 8:53 AM |
| 56 | PhD | 3/28/2018 8:26 AM |
| 57 | MSc and PhD in public health | 3/28/2018 8:24 AM |
| 58 | none | 3/28/2018 7:55 AM |
| 59 | PhD training; methodologist who conducts SR and participates in guidelines development, and has clinical training | 3/28/2018 7:47 AM |
| 60 | minimal, I am a physician specialist | 3/28/2018 7:03 AM |
| 61 | No fomal education, but have undertaken postgraduate short courses in epidemiology, various analysis techniques, systematic review and meta-analysis courses and have been working in an epidemiological department for 20 years. | 3/28/2018 6:57 AM |
| 62 | Masters and PhD in medical statistics | 3/28/2018 6:23 AM |
| 63 | registration epidemiologist B | 3/28/2018 4:06 AM |
| 64 | No degrees, just self-readings | 3/27/2018 9:41 PM |
| 65 | I'm an M.D. | 3/27/2018 8:58 PM |
| 66 | Masters degree | 3/27/2018 4:48 PM |
| 67 | I have studied epidemiology in medical school | 3/27/2018 4:47 PM |
| 68 | None | 3/27/2018 4:30 PM |
| 69 | No education in epidemiology. My education is a BScN, and a MA in Psychology. | 3/27/2018 2:49 PM |
| 70 | No formal education | 3/27/2018 2:45 PM |
| 71 | It was a course in my MPH degree | 3/27/2018 2:05 PM |
| 72 | MPH in Epidemiology | 3/27/2018 1:48 PM |

| | | |
|---|---|---|
| 73 | masters in health studies | 3/27/2018 12:47 PM |
| 74 | MSc | 3/27/2018 11:26 AM |
| 75 | MD and MPH | 3/27/2018 10:06 AM |
| 76 | No formal education. | 3/27/2018 9:29 AM |
| 77 | Masters degree in public health | 3/27/2018 8:20 AM |
| 78 | Masters level | 3/27/2018 7:16 AM |
| 79 | MSc | 3/27/2018 6:27 AM |
| 80 | Master in health economics | 3/27/2018 3:42 AM |
| 81 | Master level | 3/27/2018 3:10 AM |
| 82 | Medical school, and later personal education in relevant literature | 3/26/2018 3:35 PM |
| 83 | None | 3/26/2018 12:16 PM |
| 84 | None formal - work experience assisting with population research | 3/26/2018 11:02 AM |
| 85 | medical sociologist, undergraduate pychology, PHD in Survey design, 12 years IQWiG. | 3/26/2018 10:17 AM |
| 86 | no specific education | 3/26/2018 10:10 AM |
| 87 | Limited | 3/26/2018 9:51 AM |

## Q3 A systematic review compared the effects of cognitive behavioural therapy versus a waiting list for military suffering from post-traumatic stress disorder on depression. It found that cognitive behaviour therapy reduced depression by 8 points more on a scale from 1-100 (95% confidence interval from 21 point reduction to 12 point increase). This reduction is small but important. The evidence came from a meta-analysis with very few people (91) and very serious concern that the studies were at high risk of bias because of unclear randomisation and large loss to follow-up.The conclusion about the effect of cognitive behaviour therapy could be worded in the following three ways. Please indicate the acceptability of each statement.

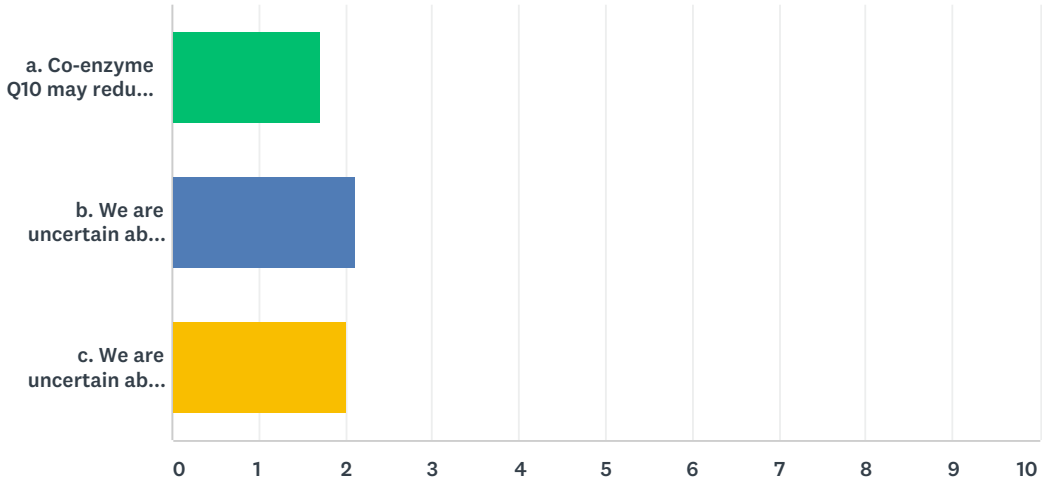
Answered: 91 Skipped: 21




| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Cognitive behaviour therapy may reduce depression slightly more than no therapy but we are uncertain. | 36.67%<br>33 | 47.78%<br>43 | 15.56%<br>14 | 90 | 1.79 |
| b. We are uncertain about the effect of cognitive behaviour therapy on depression. | 28.57%<br>26 | 51.65%<br>47 | 19.78%<br>18 | 91 | 1.91 |
| c. We are uncertain about whether cognitive behaviour therapy reduces depression more than no therapy. | 14.29%<br>13 | 60.44%<br>55 | 25.27%<br>23 | 91 | 2.11 |

# Q4 The conclusion about the effect of co-enzyme Q10 on blood pressure could be worded in the following three ways. Please indicate the acceptability of each statement.

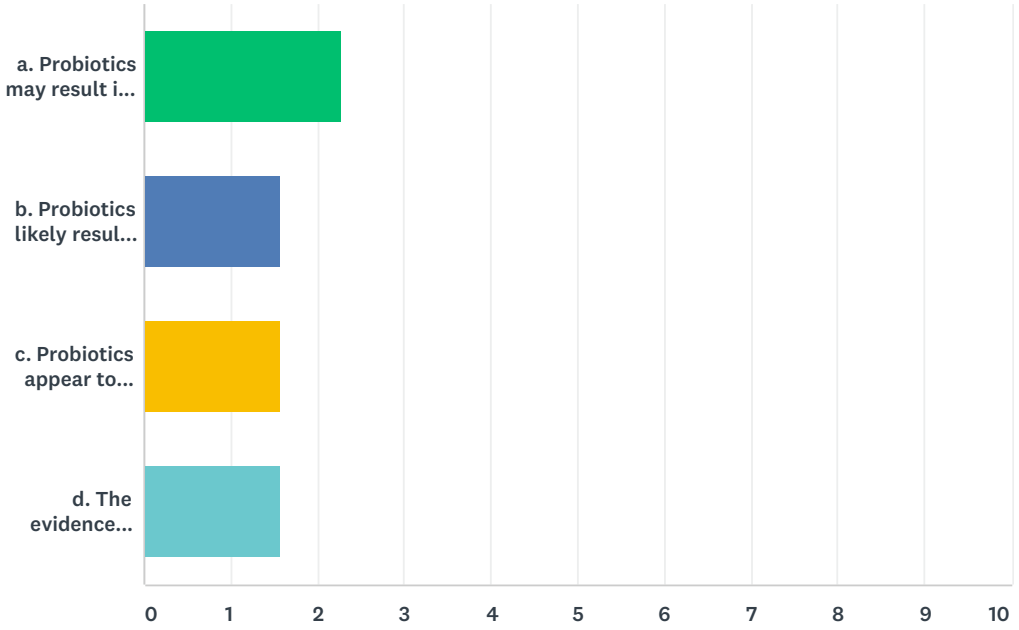
Answered: 90 Skipped: 22




| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Co-enzyme Q10 may reduce blood pressure slightly but we are uncertain. | 46.07% 41 | 35.96% 32 | 17.98% 16 | 89 | 1.72 |
| b. We are uncertain about the effect of co-enzyme Q10 on blood pressure | 15.56% 14 | 56.67% 51 | 27.78% 25 | 90 | 2.12 |
| c. We are uncertain about whether co-enzyme Q10 reduces blood pressure. | 16.67% 15 | 65.56% 59 | 17.78% 16 | 90 | 2.01 |

# Q5 The authors of the review considered that the cut-off for a large effect is RR 0.60. Please indicate the acceptability of the statements below.

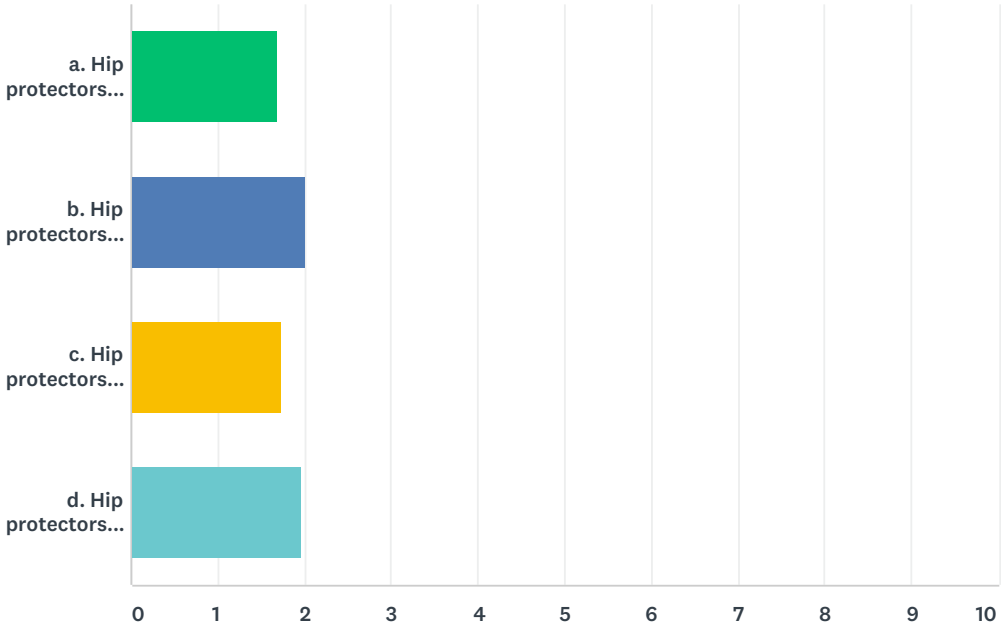
Answered: 91 Skipped: 21




| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Probiotics may result in a large reduction in the incidence of diarrhea. | 10.11% 9 | 52.81% 47 | 37.08% 33 | 89 | 2.27 |
| b. Probiotics likely result in a large reduction in the incidence of diarrhea. | 51.69% 46 | 39.33% 35 | 8.99% 8 | 89 | 1.57 |
| c. Probiotics appear to result in a large reduction in the incidence of diarrhea. | 50.00% 45 | 43.33% 39 | 6.67% 6 | 90 | 1.57 |
| d. The evidence suggests that probiotics result in a large reduction in the incidence of diarrhea. | 56.67% 51 | 31.11% 28 | 12.22% 11 | 90 | 1.56 |

# Q6 The authors indicate that the effect found was less than their cut-off for an effect. Please indicate the acceptability of the statements to communicate the effects of hip protectors compared to no hip protectors on the number of hip fractures.

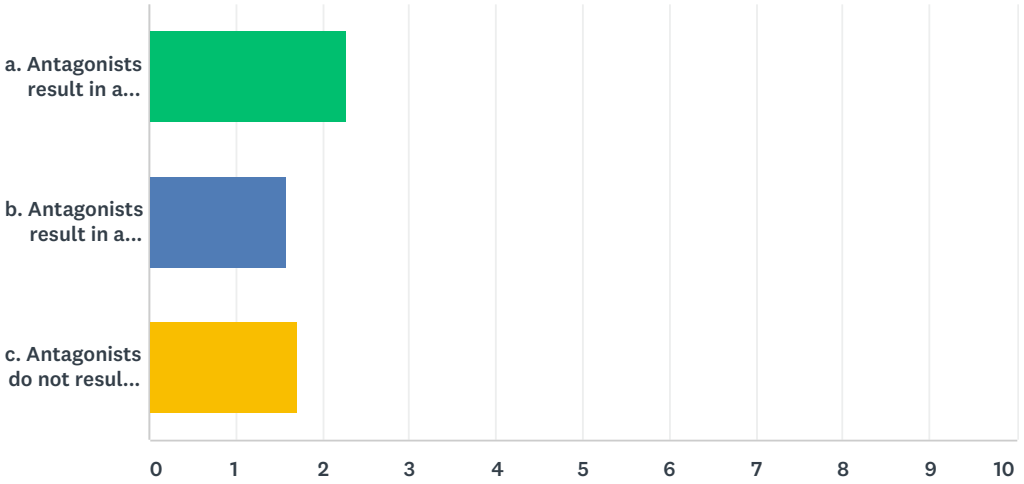
Answered: 90 Skipped: 22




| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Hip protectors likely do not reduce hip fractures. | 44.94% 40 | 40.45% 36 | 14.61% 13 | 89 | 1.70 |
| b. Hip protectors likely result in little to no difference in hip fractures. | 23.33% 21 | 52.22% 47 | 24.44% 22 | 90 | 2.01 |
| c. Hip protectors probably do not reduce hip fractures. | 39.33% 35 | 47.19% 42 | 13.48% 12 | 89 | 1.74 |
| d. Hip protectors probably result in little to no difference in hip fractures. | 21.35% 19 | 61.80% 55 | 16.85% 15 | 89 | 1.96 |

# Q7 Please indicate the acceptability of the statements below to communicate the effect of antagonists compared to placebo on daytime nasal symptoms. Note: the authors considered the SMD and confidence interval a small but not important effect.

Answered: 92 Skipped: 20




| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Antagonists result in a small effect that may not be an important reduction in daytime symptoms. | 15.22%<br>14 | 42.39%<br>39 | 42.39%<br>39 | 92 | 2.27 |
| b. Antagonists result in a small possible unimportant reduction in daytime symptoms. | 49.45%<br>45 | 41.76%<br>38 | 8.79%<br>8 | 91 | 1.59 |
| c. Antagonists do not result in an important reduction in daytime symptoms. | 45.05%<br>41 | 38.46%<br>35 | 16.48%<br>15 | 91 | 1.71 |

# Q8 Please take 5 minutes to review this list of options for statements to communicate results at various levels of evidence and size of effect. If you'd like, please provide any general comments about the statements below.

Answered: 39 Skipped: 73

| # | RESPONSES | DATE |
|---|---|---|
| 1 | Suggestion not important effect: small clinically not relevant reduction For example: High/not important: results in a small clinically not relevant reduction | 4/12/2018 3:05 PM |
| 2 | Question 5b in this survey was worded slightly different than below in a way that affected my answer: "small possible unimportant" instead of "small possibly unimportant". I think the addition of "possibly unimportant" is very helpful, but found it confusing the way it was worded in Question 5b. I would add commas: "small, possibly unimportant,..." so the grouping of these words is clear. No effect/high certainty: you can't say that something doesn't have effect. This is just wrong. Small effect important/moderate certainty: I find the word slightly to be a bit confusing. Your mind has to keep track of probably, and direction (reduces) and then amount. I would use "to some degree" instead of "slightly". "probably" works much better for me than "likely", just because it is more familiar and uses less brain power to combine the meaning of it with the other terms. | 4/6/2018 5:32 AM |
| 3 | 1. Regarding 'no effect', I think we need to differentiate clearly between two different situations which seem to currently be treated as one and described as 'no effect': 1. where there is a tight confidence interval around the null/line of no effect and the CI is completely within threshold limits - in this case, we can say that there is little to no difference in the outcome. 2. where there is a large confidence interval that crosses the null/line of no effect AND one or both important threshold limits - in this case, I think it is misleading to state that there is "little or no difference" as we cannot confirm or exclude an important difference. In this case, should we call this an 'uncertain' effect? Or that the effect is compatible with both little or no effect and an important effect? It is more difficult to describe this in plain language, but I think it is important to convey the uncertainty of the effect that exists in the second case. 2. Re the wording, "does not reduce/may not reduce" etc, I think that given that there is always some uncertainty around the confidence interval, that the word "not" is too strong. Prefer to state 'little or no difference". 3. The rest of the statements outlined below are reasonable. This is a very useful project which will help ensure correct interpretation in the results of Cochrane Reviews. Thank you! | 4/5/2018 1:08 PM |
| 4 | None of the statements specify a comparitor. I think it would be useful to ensure that the comparitor is always stated. For example, "Compared to Y, X results in a large reduction in mortality" or "X results in a large reduction in mortality compared to Y". Some of the statements include the word "evidence", while most do not. I think it would be useful to ensure that the word evidence appears in these statements to make clear they are based on evidence rather than opinion. For example, "The evidence shows that X results in a large reduction in mortality compared to Y". It is not clear to me why statements may be phrased using "probably" or "likely" and if there is a meaningful difference. I suggest picking one and discouraging use of the other. Some of the statements allude to the concept of certainty of evidence (using words like "may" and "appears"), while others make certaity explicit. I would favour making certainty explicit in all phrases. For example "The evidence shows we can be reasonably certain that X results in a large reduction in mortality compared to Y" or "The evidence shows we cannot be certain whether X reduces mortality compared to Y". The statements do not consider the fact that, while a given intervention may have little or no effect on average, it may have a large and important effect on a small number of people. They also do not consider the potential harms of an intervention. There may be interventions that are inexpensive and very safe, and which have little or no effect in most people, but do have large and important effects in some people. | 4/4/2018 4:31 AM |
| 5 | I like nuanced statements that provide precisions on which criteria are assessed (e.g. effect size, importance) and allow for uncertainty to express (e.g. likely results in little or no difference)... | 4/3/2018 10:30 PM |

| | | |
|---|---|---|
| 6 | When you think about importance of an effect size, do you consider whether a small effect could be important if it applied across a whole population? I am asking because of a recent systematic review where we and our clinician experts judged the size of the effect to be "little to no effet", but others who read our review thought the size could be important if applied at a population level | 4/3/2018 9:21 AM |
| 7 | HIGH - Instead consider "X slightly reduces mortality" (small effect (important)) - For small effect (not important) -- "X does not result in an important reduction in mortality". The direction of the observed effect is missing, and this seems important as a small effect may be meaningful to some but not all. I prefer the statements in this category that include reference to the direction of effect. MODERATE - I don't feel that "likely" and "probably" are good synonyms for MODERATE evidence. Somehow I feel that "likely" is stronger and well suited here, but "probably" is less strong and therefore not as well suited. Perhaps "likely results in" and "seems to result in" could work if you want options? LOW - I feel that "appears to" belongs in the MODERATE category and not the LOW category. I like "may" and "the evidence suggests" though. - for the "no effect" category, "X may not reduce mortality" sounds like "X may or may not reduce mortality". I find it confusing and may introduce ambiguity about being uncertain about the direction of the effect VERY LOW - I feel that some qualifier as to where the uncertainty arises (i.e. low quality evidence) is needed. What do you think about "X may reduce mortality, but the quality of evidence (or studies) is very low"? or "X may reduce mortality, but we are very uncertain due to the low quality of the evidence (or studies)" | 4/2/2018 2:43 PM |
| 8 | Wordings are fine at HIGH or MODERATE certainty of the evidence, but problems arise at LOW level of evidence. For example 'X may result in a large reduction in mortality' raises the question whether the certainty of the evidence might be higher if 'large reduction' is replaced by 'reduction' (clinically relevant reduction). In my view the conclusion should relate to what is considered a clinically relevant effect, a clinically relevant reduction (or more general: the clinical decision threshold) i.e. 'large' should not be part of the statement. The level of evidence relates to the clinical decision threshold, usually the MID, and not to a 'larger' (or smaller) effect size. In principle this is also true at HIGH or MODERATE certainty of the evidence, but because of the higher certainty of the evidence, there is less of a problem. Similarly, I think that we should be careful about conclusions on 'small (unimportant)' effects: all of our judgements should concern clinically relevant effects, statements about 'unimportant' effects are confusing and may be misleading | 4/2/2018 8:45 AM |
| 9 | Regarding the statements of 'No effect' (HIGH, MODERATE and LOW certainty): well, I have learnt that you cannot keep the null hypothesis of no effect even if you cannot reject it. So you should not say 'does not reduce mortality' or 'little or NO difference'. When the effect crosses the line of 'no effect' - the results are uncertain and inconclusive. Very low certainty: I am not sure that 'may reduce mortality but we are uncertain' is a good idea. What about saying: "The results show a reduction in mortality, but the certainty of the evidence for this is/ is assessed as very low." Also, I think in general that the expression 'we are uncertain' sounds strange. Why are 'we' uncertain? It is either the results (crossing line of no effect) or very low certainty of evidence in that case that make us uncertain - so I think we should stress that it is the results/certainty of the evidence as assessed that are uncertain - not we. We are not uncertain of that the results/evidence are uncertain: "It is uncertain whether...", "the results are uncertain becuase of the possibility of either a reduction or an increase in ..." | 4/1/2018 1:59 PM |
| 10 | I think it is important to say x results in a large reduction...COMPARED to no intervention/placebo | 3/31/2018 7:21 AM |
| 11 | Overall fewer words are preferred to communicate essential findings of quality and effect size | 3/30/2018 12:23 PM |
| 12 | Prefer likely over probably. Wording "small possibly unimportant" is very confusing to lay readers and is cognitively taxing Use of "but we are uncertain" negates "may reduce [...]; therefore no need for "may reduce" | 3/29/2018 9:25 AM |
| 13 | I am worried that the differences in wording between moderate/small important/small unimportant, with differences in the structure of sentences, use or non-use of adjective quantifiers (eg small vs slightly) and use or non-use of “important”, may lead to heterogenous interpretation between these categories. Also, as always, there are issues with translating many of these quantifier terms. | 3/29/2018 8:26 AM |
| 14 | As non-native speaker the difference between likely and probably is hard to grasp. In the previous questions I felt that 'likely' is stronger than 'probably', but that might be personal. If you look these words up in a Dutch-English dictionary the translation is largely the same. In the table I see they're both used for Moderate CiE. Same for the difference between 'appears' and 'may', these words have a different feel for me. 'Appears' is: it could be but we might be completely wrong'. 'May' sounds more certain. This is really about language, and there may be differences between countries/languages. High CiE and small not-important effect: by adding 'may not be' and 'possibly' it seems that you have doubt about the importance of the effect, but as this is High CiE that is not the case, no? | 3/29/2018 7:07 AM |

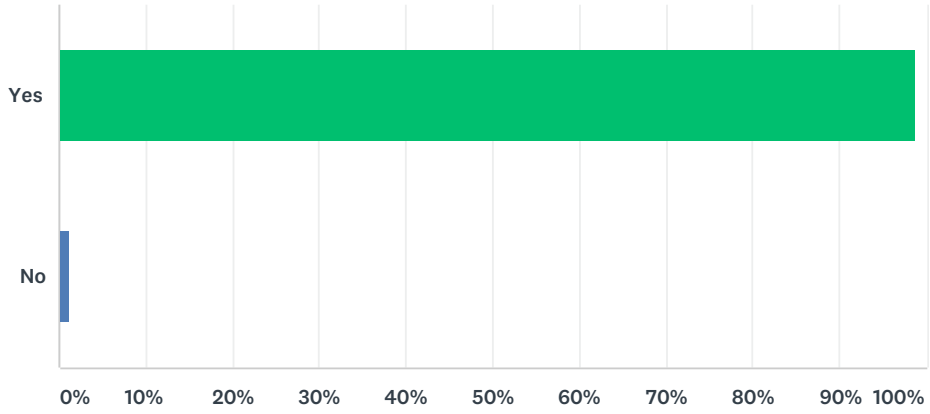
| | | |
|---|---|---|
| 15 | using the word 'appears' by LOW evidence feels too strong for me, I prefer 'may'. | 3/29/2018 6:15 AM |
| 16 | We are currently researching the communication of uncertainty in science, and use GRADE as an example of good practice because of the way it separates the different sources of uncertainty clearly - the calculated uncertainty around measurement is different from the quality of the evidence. It is also vital that the magnitude of the effect and the quality of the data are not confused (as they currently are in scales such as the IARC classifications of cancer risk, leading to statements like 'bacon in same cancer risk bracket as cigarettes' and the genomic evidence scales of 'pathogenic to benign'). I think it is important to give people as clear a view as possible of what data is currently available. However, I appreciate that for a simple statement of conclusions, such as in the preparation of guidelines, then it is important to give people a single sentence. I feel, though, that that sentence should overtly retain the distinction between quality of evidence and size of effect. So, in your examples below, I think they should be of the structure "There is strong/weak evidence that X results in a large/small effect". I don't think that there is anything to be gained in converting the phrases 'moderate evidence' to 'likely' - work on people's interpretation of words used in common parlance suggest that they lead to much more variable interpretation than keeping more precise language. | 3/29/2018 5:58 AM |
| 17 | Effects should not be communicated solely in relative terms with words like „large“, „small“ etc. Absolute effects should be added. In the probiotics example, the absolute effect is about 13 / 100. Is that large? | 3/29/2018 2:17 AM |
| 18 | The suggestion look good. I’m critical though with the choice of “slightly” for small but important effect. In your graduation of statements, it sounds too close to small unimportant instead than in between moderate an unimportant. My preference will be “small but important” difference. I have my preference on the choices provided, but I can live with all but this one. | 3/28/2018 7:24 PM |
| 19 | consistency is appreciated examples will be essential to develop consistency if these statements are for a range of audiences this should be considered | 3/28/2018 4:43 PM |
| 20 | Use appropriate comparisons for context (compared to Y); use plain language | 3/28/2018 2:06 PM |
| 21 | A very helpful tool! | 3/28/2018 1:44 PM |
| 22 | 'Appears to' and 'evidence suggests' seem too strong for low quality evidence. With those statements alone I would assume that there was certainty about the effects. Regarding the inclusion of 'important' in the interpretation - will end-users know what is meant by important? Could they use more elaboration about what is meant by an important difference? It is preferable in my opinion to include the proposed/theoretical direction of effect even if there is no effect. e.g., 'We are uncertain whether it reduces mortality' is preferable to 'we are uncertain about the effect on mortality'. May just be personal preference. | 3/28/2018 1:42 PM |
| 23 | word "appears" seems like a magical statement and not preferred in my opinion. it may be helpful to provide illustrative examples with number to contextualize some of these wording options. | 3/28/2018 12:11 PM |
| 24 | This makes sense and I agree in principle to the standardized wording. As long as you have a critical appraisal checklist whereby you can grade the certainty of the findings of the study according to very low, low, moderate and high then this system works; however not all checklists grade the quality eg CASP | 3/28/2018 9:47 AM |
| 25 | In general, the shortest text that fits the data is best (example: size of effect moderate: reduces mortality, NOT results in a reduction in mortality) if an effect is certain, small, but definitely not important, I would say: does not result in an important reduction in mortality, because clinical relevance is more important than "methodological significance". I prefer "may" over likely or probably; "suggest" is an acceptable alternative | 3/28/2018 8:16 AM |
| 26 | I like the use of certain words to communicate certain levels of certainty; however, I am unsure whether an end-user would distinguish between those words without having previously reviewed a chart like the one below -- for example, I am uncertain whether most clinicians would immediately pick up on the difference between may and likely. Also, for some reason, the word "appears" is not preferable (it may be that it isn't as commonly used in science-writing in our field; appears conjures up "magic" vs. some biologically plausible relationship) Lastly, there are some principals of plain language writing that could be integrated into some of the sample statements that would make them longer, but would also help with easier interpretation of the nuances. For example, I would prefer a statement line: X may reduce mortality; however, it is likely that the reduction is not clinically important. So, state the direction of the relationship first, and then follow with the disclaimer about clinical significance. | 3/28/2018 7:59 AM |

| | | |
|---|---|---|
| 27 | Having submitted reviewsz with this wording..most editorial boards have no knowledge of GRADe cnsider this wording vague (comments received like "is it significant or not as probably and likely or possibly is not wording we are used to see in results, please be more specific...." There is still a long way before clinicians and peer reviewers, editors, guidelin emakers are familiar with GRADE and its terminology | 3/28/2018 7:16 AM |
| 28 | Generally I support the options aligned with the certainty levels (e.g. does, results in, will for HIGH). It's not clear why some constructs are not included. For example there is no "X reduces" for HIGH-Large, and no "X resutls in" for HIGH-Small. It seems that you could have options such as "X reduces mortality substantially" and "X results in a slight reduction in mortality" for these. I find some phrases too difficult to read and understand. Particularly "X results in a small possibly unimportant effect". There are too many words next to each other! | 3/28/2018 6:52 AM |
| 29 | - I'm not a native speaker, but 'appears' seems to suggest a little more than 'low' | 3/28/2018 4:20 AM |
| 30 | I like all the statements for high and moderate certainty. I know we need to be cautious, but the statements for low certainty are so vague that we run the risk that readers have no idea what we mean. I've had feedback to this extent on recent Cochrane reviews. | 3/27/2018 5:28 PM |
| 31 | I think it is a great idea to use such tool to standardize the way SR results are interpreted. | 3/27/2018 4:52 PM |
| 32 | The double negative of some statements can be confusing. For example: X likely results in a small effect that may not be an important reduction in mortality. I find for someone who is not well versed in research, that can cause confusion. However, many of the other statements are clear. | 3/27/2018 2:58 PM |
| 33 | I prefer the use of "probably" versus "likely" | 3/27/2018 2:11 PM |
| 34 | Language preference is for: HIGH: "small possible unimportant effect in mortality" rather than "does not results in an important reduction in mortality" and "results in little to no difference in mortality" rather than "does not reduce mortality". MODERATE: "likely" rather than "probably". LOW: "may result" rather than "appears to" | 3/27/2018 9:51 AM |
| 35 | Statements need to be simple. "X results in a small effect that may not be an important reduction in mortality" is too complicated a statement, for example - not to mention the equivalent statements for lower QoE. General comments: - For a critical binary outcome, especially mortality, is there a such a thing as an unimportant effect? How small would that be? - Why distinguish between an unimportant effect and no effect? If this is to do with the problem of averages and individual variance for continuous outcome measures (e.g. there is no important effect on average but some patients may have a substantial benefit/Response), this should be made explicit. | 3/27/2018 8:55 AM |
| 36 | I like to work with the standard phrasing as much as possible.It is great for abstracts and is as good as any way to integrate the size and certainty of effect. There are two situations I find troubling to work with: 1. Moderate quality evidence of moderate effect when you have downgraded for imprecision. I always want to describe this in terms of greater uncertainty than 'probably' or even 'may'. I actually end up working something like: 'We found moderate quality evidence that [intervention] reduces mortality by about 4% when compared with control, although we do not have enough data from the studies to rule out there being little or no effect/potential increase in risk of death (18% versus 14%; RR 0.73 (95% CI 0.34 to 1.2)' 2. Very low quality evidence and large/moderate effect when you have a large amount of data from the analyses: SMD 0.5 (0.2 to 0.8; 29 studies, 2300 participants). I don't always think that 'There is insufficient evidence to determine the effects of' is really right here so end up with something like: 'Our confidence in the moderate reduction in symptoms with intervention is very low due to bias and variation between the results of the studies'. | 3/27/2018 7:17 AM |
| 37 | 1)X reduces mortality slightly - why not: reduction in mortality was small 2)"small possibly unimportant" - could be confusing to people. It could be unclear to people what does it mean "possibly unimportant" 3) When talking about "important reduction", some context is necessary. What is important reduction? In the context of this information, it becomes apparent what is important or unimportant reduction 4) "appears to result" has potential to be misleading. The results should be described with more clarity. 5) There is not much difference in appears/may - the wording is not very clear | 3/26/2018 3:48 PM |
| 38 | The table is too complex to retain for mostly SR readers (i.e. non experts about SR production, like primary care physicians...). | 3/26/2018 12:42 PM |

| | | |
|---|---|---|
| 39 | Does "importance" always have to do with what is already known about the disease? Should every reported outcome have a known level of importance? At a certain threshold, do all effect sizes become "important" or can a moderate effect be "not important"? What exactly determines the difference between small important/small not important? It would be helpful to see this clarified with effect sizes. | 3/26/2018 12:00 PM |

# Q10 One last question: Do you agree in principle that conclusions should be based on the concepts of the importance/size of the effect and the certainty of the evidence?

Answered: 85 Skipped: 27



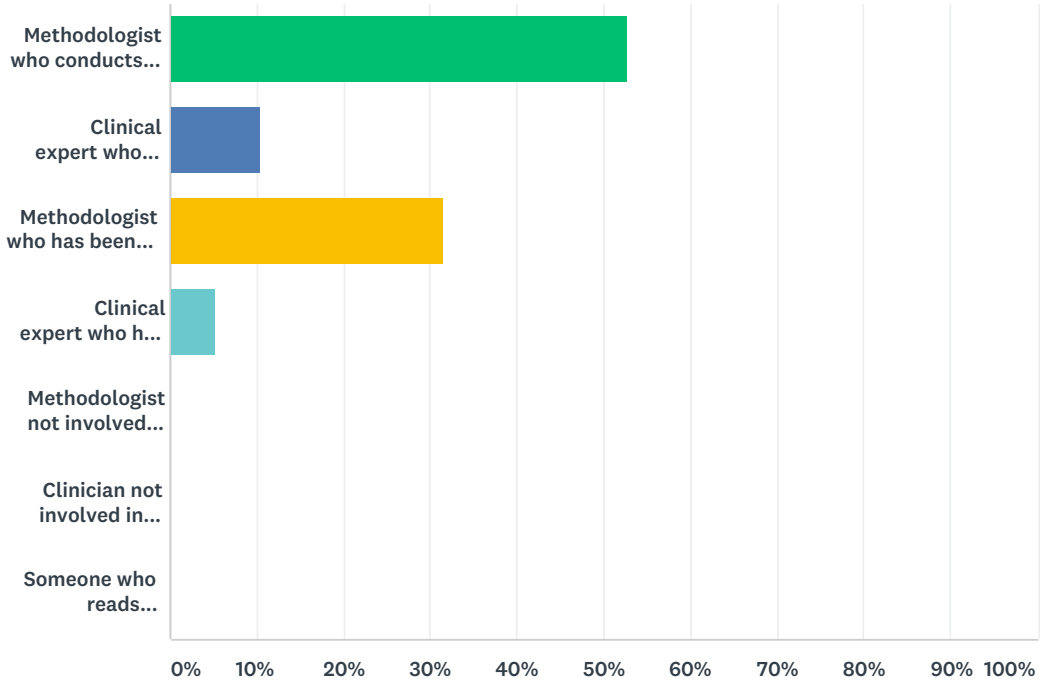

| ANSWER CHOICES | RESPONSES | |
|---|---|---|
| Yes | 98.82% | 84 |
| No | 1.18% | 1 |
| TOTAL | | 85 |

| # | PLEASE PROVIDE COMMENTS IF YOU'D LIKE. | DATE |
|---|---|---|
| 1 | good idea! | 4/6/2018 5:32 AM |
| 2 | I do agree with this statement, but I think there is a risk that importance and effect size may be conflated into a single concept. To me, importance includes aspects such as quality of life, possible adverse events, and patient preferences, while effect size is a necessarily specific quantitative measure. I reiterate what I said above about the problem with excluding information about potential harms (perhaps subsumed into the conceopt of importance), the fact that an intervention that has little or no effect on average could actually have a large and important effect in some people. I think it is also unwise to ignore practical issues such as the cost of an intervention (or cost relative to an alternative) or how difficult it would be to implement the intervention in the target population. | 4/4/2018 4:31 AM |
| 3 | Totally agree with both concepts needing to be included in the conclusion -- I only feel that we may need to explain WHY we are uncertain (i.e. low quality of evidence) as opposed to just saying we're uncertain (for the VERY LOW quality category). Left unqualified "uncertainty" can stem from many situations, so I think clearer to say because of low quality evidence. | 4/2/2018 2:43 PM |
| 4 | Yes and NO: see my earlier comments, I think we should not be making statements on 'unimportant' (i.e. not clinically relevant/ not patient relevant) effects. Also there is a fundamental issue: the level of evidence relates to a clinically relevant effect (or similar clinical decision threshold) and not to a 'larger' (or smaller) effect. For example with a grading HIGH we might be very certain that a particular intervention is superior i.e. has a clinically relevant advantage as compared to the control, but how certain would we be that the advantage is 'large'? In theory one could define 'large' and use this as a threshold to judge e.g. imprecision, and come up with a grading for a 'large effect' (in my example the level of evidence could be HIGH for a clinically relevant effect, and MODERATE for a 'large effect'), but this would make the GRADE system much too complicated and confusing! | 4/2/2018 8:45 AM |
| 5 | However, I must also say that as a methodologist and not a practitioner, it is often difficult to assess the size/importance of an effect. We have been told by our reviewers to leave that judgement to the practitioners. | 4/1/2018 1:59 PM |

| | | |
|---|---|---|
| 6 | I think certainty of evidence is an improvement over the concept of quality of evidence. When it comes to small effects, I am not sure I understand how we will decide between important and not important. | 3/30/2018 5:55 PM |
| 7 | results of this survey will be helpful to guideline developers. | 3/30/2018 12:23 PM |
| 8 | What I find difficult is that the wording is about the importance/effect size and CiE, but what I (and many colleagues) automatically do when interpreting the results is to take the CI into account (wide CI make you less certain). I know this is not GRADE guidance, and I see that it makes sense conceptually, but to me it's still counter-intuitive. | 3/29/2018 7:07 AM |
| 9 | Yes, it's vital that guidelines and conclusions reflect both (and indeed the clinical importance rather than the statistical significance of results). It does not mean that we have to wait for strong evidence to be available before doing anything, but the quality of evidence should be taken into account when making decisions. | 3/29/2018 5:58 AM |
| 10 | There are other concepts that should be incorporated | 3/28/2018 2:06 PM |
| 11 | A- methodologists' opinion should not matter when we are trying to standardise the readability for end user, so I think surveying the likes of us is not going to get useful answers. B- A large chunk of SR work and its use is happening in English as a second language people and countries (Think Europe even). There is very little difference between the statements that I just saw from that perspective (I have dealt directly with some decision makers and stakeholders in non English countries). Standardising on our preferences and terms 'likely/ probably/ possibly' won't improve readability for the end user if they are not native English speakers. C- I longed for an open ended answer option for each question throughout, because I don't word a result statement without referring to the whole PICO - personal preference. | 3/28/2018 8:44 AM |
| 12 | Separating size of effect and certainty would be a helpful advance - similar to the separation of QoE and recommendations - especially if judgements/decisions were to be made transparent. | 3/27/2018 8:55 AM |
| 13 | But the context is very important, to explain to people why something is important or not important. | 3/26/2018 3:48 PM |
| 14 | Sure. "(significance + grade of evidence) + (size + direction) of the effect " | 3/26/2018 12:42 PM |
| 15 | Unless patients are not regularly involved in choosing / weighing endpoints for systematic reviews the concept of importace may differ between clinicians / authors and patients | 3/26/2018 10:20 AM |

# Q1 What is your primary role related to systematic reviews and guidelines? (Select one that best represents your role)

Answered: 19 Skipped: 0




| ANSWER CHOICES | RESPONSES | |
|---|---|---|
| Methodologist who conducts systematic reviews | 52.63% | 10 |
| Clinical expert who conducts systematic reviews | 10.53% | 2 |
| Methodologist who has been involved in guideline development | 31.58% | 6 |
| Clinical expert who has been involved in guideline development | 5.26% | 1 |
| Methodologist not involved in systematic reviews or guidelines | 0.00% | 0 |
| Clinician not involved in systematic reviews or guidelines | 0.00% | 0 |
| Someone who reads systematic reviews | 0.00% | 0 |
| TOTAL | | 19 |

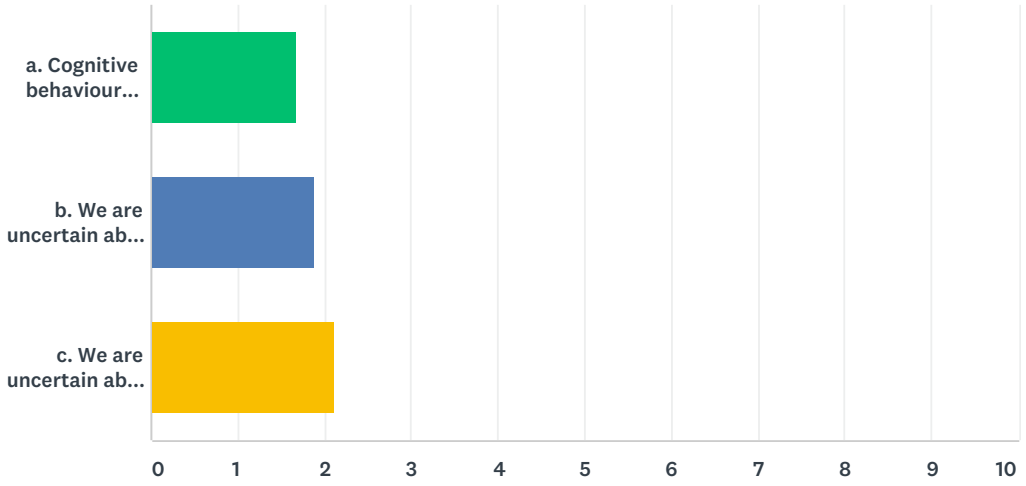
# Q2 What is your education in epidemiology?

Answered: 17 Skipped: 2

| # | RESPONSES | DATE |
|---|---|---|
| 1 | PhD | 4/3/2018 3:15 AM |
| 2 | Part of my undergraduate training. 'On-the-job' training undertaking systematic reviews and teaching people how to undertake systematic reviews | 4/2/2018 8:07 PM |
| 3 | Masters degree | 3/30/2018 5:25 AM |
| 4 | None. | 3/28/2018 12:56 AM |
| 5 | PhD in health research methods and MPH | 3/27/2018 3:42 PM |
| 6 | Epidemiologist responsible for postgraduate teaching and supervision of epidemiology | 3/27/2018 10:09 AM |
| 7 | nutritional epidemiology | 3/27/2018 9:38 AM |
| 8 | equivalent to a masters | 3/27/2018 8:55 AM |
| 9 | PhD(c) | 3/26/2018 8:36 PM |
| 10 | I attended workshops | 3/26/2018 3:33 PM |
| 11 | PhD Degree | 3/26/2018 1:32 PM |
| 12 | PhD | 3/26/2018 11:34 AM |
| 13 | No formal education in epidemiology. PhD in experimental medicine. | 3/26/2018 10:08 AM |
| 14 | Master degree in Epidemiology | 3/26/2018 9:45 AM |
| 15 | Mph- masters, epidemiology | 3/26/2018 7:32 AM |
| 16 | Quantitative and qualitative meta-analysis methods. NMA training in WinBugs. | 3/26/2018 6:58 AM |
| 17 | PhD | 3/26/2018 6:37 AM |

# Q4 A systematic review compared the effects of cognitive behavioural therapy versus a waiting list for military suffering from post-traumatic stress disorder on depression. It found that cognitive behaviour therapy reduced depression by 8 points more on a scale from 1-100 (95% confidence interval from 21 point reduction to 12 point increase). This reduction is small but important. The evidence came from a meta-analysis with very few people (91) and very serious concern that the studies were at high risk of bias because of unclear randomisation and large loss to follow-up.The conclusion about the effect of cognitive behaviour therapy could be worded in the following three ways. Please indicate the acceptability of each statement.

Answered: 19 Skipped: 0



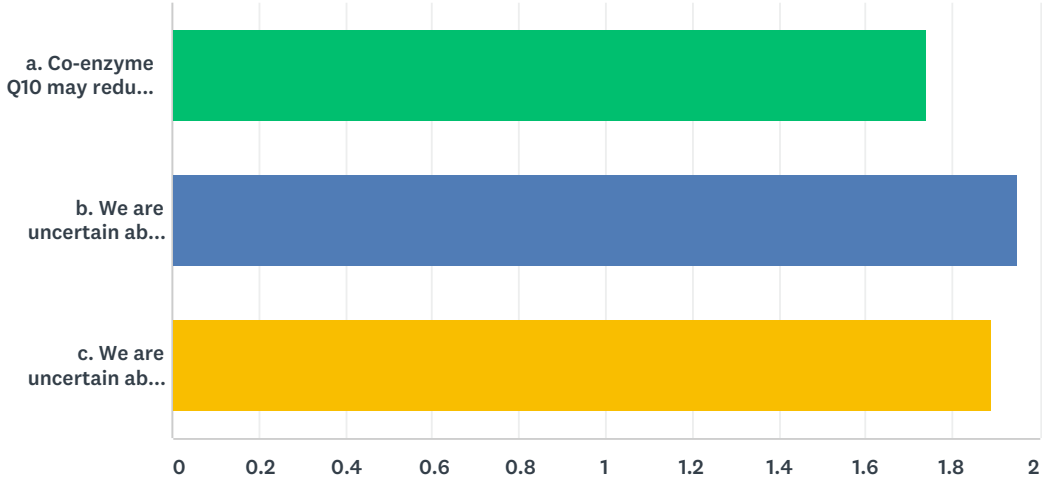

| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Cognitive behaviour therapy may reduce depression slightly more than no therapy but we are uncertain. | 47.37% 9 | 36.84% 7 | 15.79% 3 | 19 | 1.68 |
| b. We are uncertain about the effect of cognitive behaviour therapy on depression. | 36.84% 7 | 36.84% 7 | 26.32% 5 | 19 | 1.89 |
| c. We are uncertain about whether cognitive behaviour therapy reduces depression more than no therapy. | 26.32% 5 | 36.84% 7 | 36.84% 7 | 19 | 2.11 |

| # | PLEASE PROVIDE COMMENTS IF YOU HAVE ANY. | DATE |
|---|---|---|
| 1 | Difference between b and c depends if it is in the context of an SoF-table. If yes, then I prefer b. if not, I prefer c. | 4/3/2018 3:19 AM |
| 2 | So this is a low certainty of effect situation which is what makes the example hard. I prefer to give a hint to what the effect estimate is even with low certainty evidence. | 3/27/2018 3:45 PM |
| 3 | I assume that the certainty in this example is very low (downgraded twice for risk of bias and once for imprecision). While I appreciate the logic of not indicating any effect direction within the conclusion, I think this is problematic for many/most(?) reviews. In my areas of research, most assessments come out as very low with the occasional low. | 3/27/2018 10:27 AM |

| | | |
|---|---|---|
| 4 | Even in case of very low certainty of the evidence I think you should state an effect was observed. | 3/27/2018 9:44 AM |
| 5 | I would find the following acceptable: Cognitive behaviour therapy may reduce depression slightly more than no therapy but we are very uncertain. | 3/26/2018 8:38 PM |
| 6 | a. The uncertainty is reflected in the "may". The "but we are uncertain" part is redundant and may seem contradictory and result in confusion. It does depend, however, on how much uncertainty you believe is expressed in the word "may" c. Acceptable but too many words. Readers may spend a lot of time trying to make sure they understand the statement correctly, and when they finally do they realize that there is no answer to the question | 3/26/2018 1:38 PM |
| 7 | I would have rated this as very low certainty and would rather say "it is unclear whether cognitive..." | 3/26/2018 10:20 AM |

# Q5 The conclusion about the effect of co-enzyme Q10 on blood pressure could be worded in the following three ways. Please indicate the acceptability of each statement.

Answered: 19 Skipped: 0

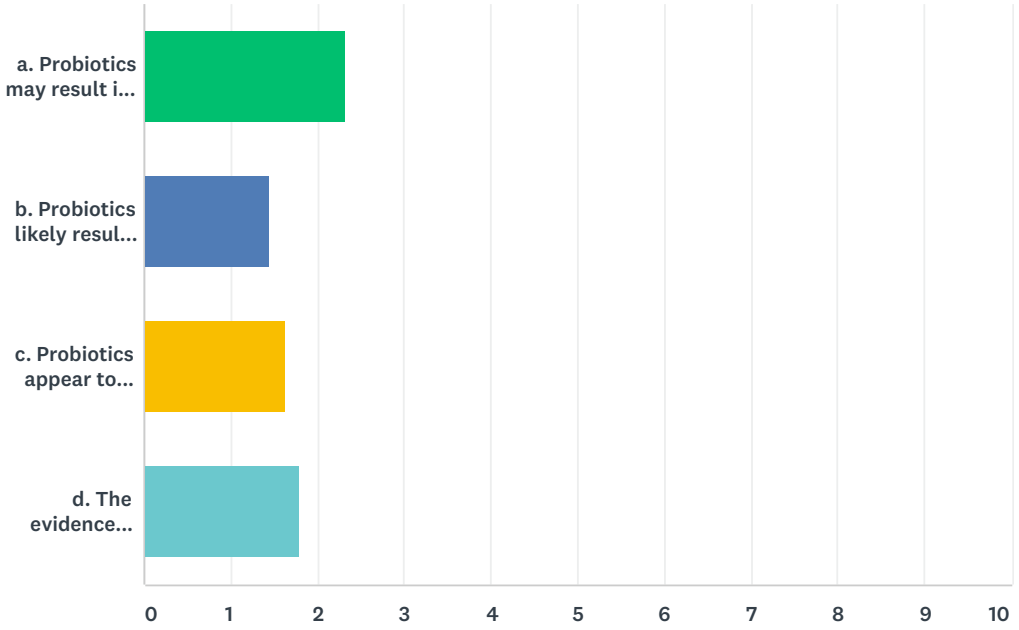



| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Co-enzyme Q10 may reduce blood pressure slightly but we are uncertain. | 52.63% 10 | 21.05% 4 | 26.32% 5 | 19 | 1.74 |
| b. We are uncertain about the effect of co-enzyme Q10 on blood pressure | 36.84% 7 | 31.58% 6 | 31.58% 6 | 19 | 1.95 |
| c. We are uncertain about whether co-enzyme Q10 reduces blood pressure. | 26.32% 5 | 57.89% 11 | 15.79% 3 | 19 | 1.89 |

| # | PLEASE PROVIDE COMMENTS IF YOU HAVE ANY. | DATE |
|---|---|---|
| 1 | This example also adds the issue of small effect and the implication from a population point of view. so you may want to say "reduces individual's BP" | 3/27/2018 3:48 PM |
| 2 | the rationale is the same as I stated before | 3/27/2018 9:45 AM |
| 3 | I would find the following acceptable: Co-enzyme Q10 may reduce blood pressure slightly but we are very uncertain. | 3/26/2018 8:39 PM |
| 4 | is it 1.6 the difference between both arms or the absolute effect? I assume the difference is meant, but very misleading in the SOF! Lower by xxx more is quite confusing... In all the statements compared to placebo is missing | 3/26/2018 3:45 PM |
| 5 | a. Same comment as previous question For b vs c, it's a matter of what you are certain or uncertain of. I would say it makes little difference when the certainty is very low, but it may be more important when you have moderate certainty due to imprecision, and you have to communicate whether you have moderate certainty of an effect or of a lack of effect | 3/26/2018 1:46 PM |
| 6 | I would use may if low certainty. Again would rather use unclear, and in that case c) | 3/26/2018 10:22 AM |

# Q6 The authors of the review considered that the cut-off for a large effect is RR 0.60. Please indicate the acceptability of the statements below.

Answered: 19 Skipped: 0



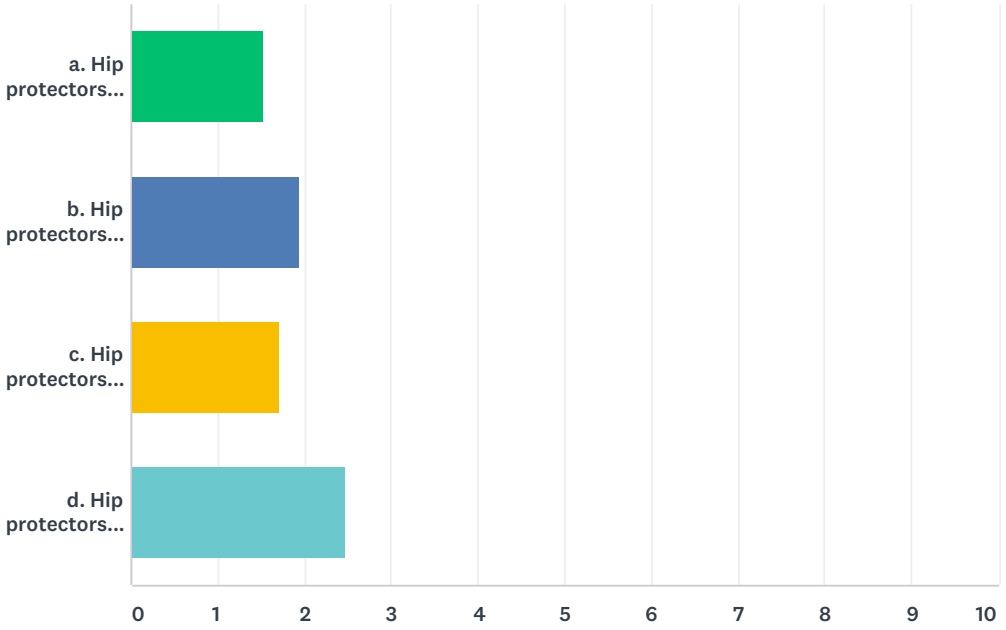

| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Probiotics may result in a large reduction in the incidence of diarrhea. | 15.79% 3 | 36.84% 7 | 47.37% 9 | 19 | 2.32 |
| b. Probiotics likely result in a large reduction in the incidence of diarrhea. | 66.67% 12 | 22.22% 4 | 11.11% 2 | 18 | 1.44 |
| c. Probiotics appear to result in a large reduction in the incidence of diarrhea. | 42.11% 8 | 52.63% 10 | 5.26% 1 | 19 | 1.63 |
| d. The evidence suggests that probiotics result in a large reduction in the incidence of diarrhea. | 42.11% 8 | 36.84% 7 | 21.05% 4 | 19 | 1.79 |

| # | PLEASE PROVIDE COMMENTS IF YOU HAVE ANY. | DATE |
|---|---|---|
| 1 | b&c: related to moderate QoE to me. d: I wouldn't move into this direction, the word 'suggest' might be confused with terminology for weak recs. | 4/3/2018 3:30 AM |
| 2 | It will be very concerning if we ignore the certainty in the statements. I prefer to explicitly state that we are not certain in addition to the "may" or "appear to" | 3/27/2018 3:51 PM |
| 3 | Given there is low certainty, that should ideally be communicated within the conclusion. The word 'suggests' seems to communicate that best but all seem acceptable. | 3/27/2018 10:36 AM |
| 4 | b is unacceptable: "likely" is in contradiction with low QoE. | 3/27/2018 9:48 AM |
| 5 | The wording needs to include some reflection that the evidence is of low certainty. | 3/26/2018 8:46 PM |
| 6 | as an adjunct to antibiotics compared to placebo... | 3/26/2018 3:49 PM |
| 7 | Maybe it's for simplicity for this survey but I wonder if the guidance should be that the narrative statements regarding magnitude should be made based on the absolute effects. May, appear, and suggest all communicate considerable uncertainty to me, which matches the low certainty. Likely sounds too strong. | 3/26/2018 1:51 PM |

| | | |
|---|---|---|
| 8 | Option d could be confused with a weak recommendation | 3/26/2018 1:48 PM |
| 9 | I think the word large should be omitted from all statements, since 0.6 is outside the 95% ci. | 3/26/2018 7:39 AM |

# Q7 The authors indicate that the effect found was less than their cut-off for an effect. Please indicate the acceptability of the statements to communicate the effects of hip protectors compared to no hip protectors on the number of hip fractures.

Answered: 17 Skipped: 2



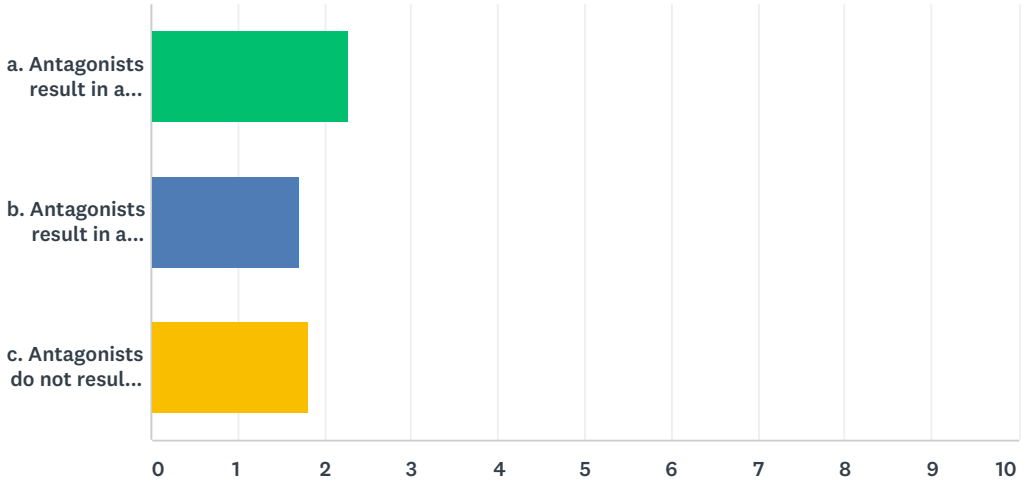

| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Hip protectors likely do not reduce hip fractures. | 58.82% 10 | 29.41% 5 | 11.76% 2 | 17 | 1.53 |
| b. Hip protectors likely result in little to no difference in hip fractures. | 25.00% 4 | 56.25% 9 | 18.75% 3 | 16 | 1.94 |
| c. Hip protectors probably do not reduce hip fractures. | 47.06% 8 | 35.29% 6 | 17.65% 3 | 17 | 1.71 |
| d. Hip protectors probably result in little to no difference in hip fractures. | 11.76% 2 | 29.41% 5 | 58.82% 10 | 17 | 2.47 |

| # | PLEASE PROVIDE COMMENTS IF YOU HAVE ANY. | DATE |
|---|---|---|
| 1 | b&d: ideal (if you feel probably and likely are interchangeable, I do) a&c: I don't like 'do not', 'little to no difference' solves a lot of problems. | 4/3/2018 3:38 AM |
| 2 | The use of likely is important because it reflects the certainty | 3/27/2018 3:52 PM |
| 3 | The clinical question is about reducing hip fracture risk so communicating the take-home message in relation to that makes sense. The 'little to no difference' is obviously technically correct but less intuitive and understandable. | 3/27/2018 10:44 AM |
| 4 | b/d reflects there was an effect although (very) small. Note: In Dutch language it is hard to know what is the difference between likely and probably. | 3/27/2018 9:55 AM |

| | | |
|---|---|---|
| 5 | "The authors indicate that the effect found was less than their cut-off for an effect." - do you mean less than cut off for an important effect? Assuming that the CI interval excludes an important difference, then b or d is acceptable. I feel strongly that we should never ever rate our certainty that there is 'no effect' -- there will always be infinitely too much imprecision. | 3/26/2018 8:52 PM |
| 6 | This one is hard. Some things to consider - The RR seems to be for an increase in hip fractures (there are more fractures in the HP group), so why are all the statements about HP reducing the risk? To me this is about providing guidance with regards to how to frame the statement: based on what we expected to see (in this case, HP and their likelihood of reducing fractures), or what we saw (in this case, HP increased fractures). - If the former, should the statement about the presence of an effect versus the presence of no effect be based on/ modified by what we expected to see? - I said this before, but it may be important to be explicit about whether the statement is based on the point estimate, the CI, or when is which. In this example it you use the CI (no effect vs little to no effect) while in the others you used the point estimate+ CI (effect vs no effect), even if you had less certainty. | 3/26/2018 2:04 PM |

# Q8 Please indicate the acceptability of the statements below to communicate the effect of antagonists compared to placebo on daytime nasal symptoms. Note: the authors considered the SMD and confidence interval a small but not important effect.

Answered: 17 Skipped: 2




| | UNACCEPTABLE | ACCEPTABLE | IDEAL | TOTAL | WEIGHTED AVERAGE |
|---|---|---|---|---|---|
| a. Antagonists result in a small effect that may not be an important reduction in daytime symptoms. | 11.76% 2 | 47.06% 8 | 41.18% 7 | 17 | 2.29 |
| b. Antagonists result in a small possible unimportant reduction in daytime symptoms. | 35.29% 6 | 58.82% 10 | 5.88% 1 | 17 | 1.71 |
| c. Antagonists do not result in an important reduction in daytime symptoms. | 47.06% 8 | 23.53% 4 | 29.41% 5 | 17 | 1.82 |

| # | PLEASE PROVIDE COMMENTS IF YOU HAVE ANY. | DATE |
|---|---|---|
| 1 | b: I would not try to incorporate any uncertainty about the clinical threshold within this standard statement, I feel cognitively that really is too much, and too confusing. We set a treshold (and that is of course always uncertain and highly depending on a lot of factors, but nevertheless it complicates these statements if we try to incorporate this). I would reserve the word 'possibly' as one of the option for expressing uncertainty in case of QoE LOW. | 4/3/2018 3:51 AM |
| 2 | If the reduction is not of any clinical importance, mentioning it within the conclusion is misleading and muddies the water. The wording "may not be an important reduction" and "possible unimportant reduction" are a bit strange. Surely, there is little uncertainty in the decision about whether these are or are not important (unless it is genuinely on the borderline of clinical importance)? | 3/27/2018 10:57 AM |
| 3 | b is more concise than a. Thatis why I prefer this one. c lacks nuance. | 3/27/2018 9:57 AM |

| | | |
|---|---|---|
| 4 | To say with any certainty that the effect is unimportant, then we need to also know the distribution of the effect. Is it normal? For example, are there some people in whom antagonists confer a large important effect and in others no effect? Have a look at how we took into consideration the distribution in this guideline: https://www.bmj.com/content/357/bmj.j1982 Ideal would be: Antagonists result in a small reduction in daytime symptoms that most people would probably not consider important. We also need to have a high degree of certainty in the MID to say anything about importance, don't we? We also need to be certain (or at least make the assumption) that most people have similar MIDs, or at least that most people would not consider it important. I'm not sure that's the case here... So I think really what we are rating is that there is a difference. Probably better to say that there is a difference and leave it at that, unless you can provide more information as above. | 3/26/2018 9:03 PM |
| 5 | a. OK but too wordy and confusing b. OK but reads weird to me. Maybe "small but unimportant", "small but not important" c. Does not present the full picture in which you have high certainty. It could mean that there is a small and unimportant effect or that there is no evidence of effect (not statistically significant) | 3/26/2018 2:07 PM |
| 6 | I prefer c) but important to clearly define the chosen threshold for an important effect | 3/26/2018 10:32 AM |

# Q9 Please take 5 minutes to review this list of options for statements to communicate results at various levels of evidence and size of effect. Please add any general comments about the statements below.

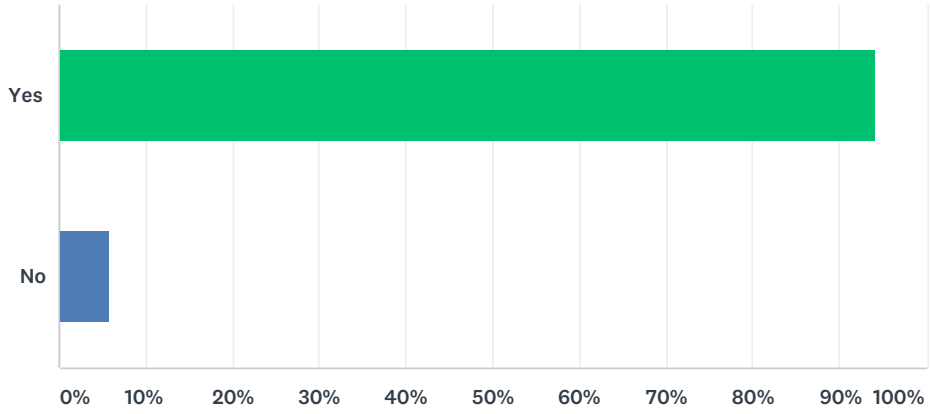
Answered: 12 Skipped: 7

| # | RESPONSES | DATE |
|---|---|---|
| 1 | Some of the statements are very clumsy with poor grammar, hard to understand what they mean without reading them several times | 4/3/2018 9:41 AM |
| 2 | In a lot of situations 'little to no difference' solves a lot of issues because panelist in general seem to have problems when we say there is no effect when we the point estimate suggest there is (but our judgements say there is a small effect that is rated as not important to patients). Major issue is using standardized statements in the context of a SoF. Within the context of a SoF: my preference is to keep them as brief as possible (but giving just enough information readers are able to understand (by for example making the assumption that the comparison is clear form the table: so no need to repeat in the statement), and to be able to present them in user friendly format.) As soon as people decide to go outside the SoF readability usually really decreases, because they often feel the urge to incorporate all the info that is in a SoF:) Last point: I feel this colum in SoF is really crucial, because - if we like it or not - most end-users (naive or not) need help interpreting what the SoF says per outcome. People also use the word 'possibly' indicating LOW quality evidence. Might be another option. I would leave out the option 'X probably may result in a small possibly unimportant effect'. Very simply put: I currently think of these statments as expressing 1) the uncertainty (is/probably is/possibly is/uncertain) and 2) judgement about the magnitude (call it clinically relevance or importance to patient) preferably based on cutt-of value. I used the options below quiet a lot (and we translated them into Dutch), and in almost all instances they work for me. | 4/3/2018 4:19 AM |
| 3 | Think the 'small possibly unimportant' statement is confusing and prefer the 'small effect that may not be an important reduction'. The 'does not result in an important reduction' seems too blunt. The 'moderate' category doesn't have an adjective to it whereas the large and small categories do Prefer 'likely' as opposed to 'probably' Why in the Low certainty of evidence section is there an option of 'The evidence suggests' when this is not in the other sections? Very low section - prefer the 2nd and 3rd statements as opposed to the 1st | 4/2/2018 9:14 PM |
| 4 | I tend to prefer the terms probably for moderate certainty and may for low certainty. Likely and appears also seem to be acceptable synonyms. I would tend not to use may for very low certainty evidence. | 3/30/2018 5:49 AM |
| 5 | Did you consider adding the following to the list We are certain, x results in large reduction We are mod certain, x results in large effect We are uncertain, x results in large effect | 3/27/2018 4:01 PM |
| 6 | For small not important effect, I prefer the bottom of the three suggested options. I also generally prefer the top option for no effect. For moderate certainty and a small effect (not important), I think it would be better to replace "may not be an important reduction" with "is not an important reduction". The wording of "small possibly unimportant effect" could be replaced by "small but unimportant effect". | 3/27/2018 11:12 AM |
| 7 | Wordings are generally quite acceptable. But in case of very low certainty I would like to suggest to add "very" to uncertain in the last two sentences. | 3/27/2018 10:04 AM |
| 8 | Making mention of a magnitude of effect requires a values judgement. One that can be empirically measured in a population. If the authors use their own judgement (always extremely dubious), then they need to be explicit. I don't think that we can ever have high or moderate certainty that an effect is large or small without also having certainty that all or almost all people would weigh the magnitude in effect similarly. There is probably a huge amount a variability between people in how much value they place on different effect sizes. At least that is my experience. So, if we include a magnitude of effect in the certainty ratings, we need to incorporate our certainty in the typical and distribution of values and preferences of those who the evidence applies to. | 3/26/2018 9:11 PM |

| | | |
|---|---|---|
| 9 | High certainty: small effect (not important): x results in a small possibly unimportant sounds very confusing for the reader, I suggest to remove this option same is true for moderate certainty, even more confusing: probably small possibly unimportant. What does this mean?? As I am not a native speaker: Is probably and likely exactly the same in English? low certainty: is may and appear exactly the same? For me, may sounds more convincing...I prefer slightly over may and appears for low certainty. For no effect, low certainty I prefer the sentences with the "little to no difference" very low certainty: I prefer the sentences starting with: we are uncertain | 3/26/2018 4:02 PM |
| 10 | It is not clear to me whether "The evidence suggests...." is a necessary option in case of low evidenice. It is not used in other scenarios at all. I could imagine that this makes a difference in terms of commuicating results and/or conclusions. The statement "the evidence suggests..." is quite unpersonal. This might also imply that evidence suggests XYZ, but we (the authors) think that ABC. All other statements are more straight forward in this sense, in my opinion. | 3/26/2018 10:00 AM |
| 11 | This is a good start that needs some discussion going forward. | 3/26/2018 7:46 AM |
| 12 | Although stronger statements are reasonable when certainty is higher, the imprecision is already factored in. So the effect size needs to be considered in terms of absolute effects. | 3/26/2018 7:07 AM |

# Q10 One last question: Do you agree in principle that conclusions should be based on the concepts of the importance/size of the effect and the certainty of the evidence?

Answered: 17 Skipped: 2




| ANSWER CHOICES | RESPONSES | |
|---|---|---|
| Yes | 94.12% | 16 |
| No | 5.88% | 1 |
| TOTAL | | 17 |

| # | PLEASE PROVIDE COMMENTS IF YOU'D LIKE. | DATE |
|---|---|---|
| 1 | Also have to consider impact of adverse effects and safety | 4/3/2018 9:41 AM |
| 2 | I fully agree! And this is really a great help formulating conclusion /evidence statements. Thanks for this great work. | 4/3/2018 4:19 AM |
| 3 | This is extremely important to highlight . The certainty of the evidence is what usually gets lost | 3/27/2018 4:01 PM |
| 4 | The importance/size of the effect is crucial and often not given adequate consideration. | 3/27/2018 11:12 AM |
| 5 | The effect size is very important. It will help you in balancing benefits and harms (in the EtD framework) | 3/27/2018 10:04 AM |
| 6 | I don't think we're there yet. | 3/26/2018 9:11 PM |
| 7 | Yes, but these should always be in relation to each other, ie the certainty rating should be made in relation to the defined range of importance/size of effect. | 3/26/2018 10:36 AM |

## APPENDIX 3: Examples of using informative statements

### Example 1

| | |
|---|---|
| Review: | Gibson M et al. Welfare-to-work interventions and their effects on the mental and physical health of lone parents and their children. Cochrane Database of Systematic Reviews 2018, Issue 2. Art. No.: CD009820. |
| Intervention: | Welfare-to-work interventions to improve employment and income in lone parents compared to usual care |
| Outcome: | Maternal health/Number of mothers reporting poor health |
| Risk Ratio: | 0.85 (95% CI, 0.54 to 1.36) |
| Absolute effects: | 30 per 1000 fewer mothers reporting poor health (from 92 fewer to 73 more)<br><br>If the authors set a threshold of 40 fewer mothers reporting poor health as a small important effect, the point estimate is identified as a **trivial, small unimportant effect or no effect.** |
| Certainty of evidence: | **Moderate certainty** (rated down once for imprecision due to wide confidence intervals including a small important effect) |

**Options for informative statements**

Welfare-to-work interventions probably results in little to no effect on maternal health

Welfare-to-work interventions likely results in little to no effect on maternal health

Welfare-to-work interventions probably does not improve maternal health

Welfare-to-work interventions likely does not improve maternal health

### Example 2

**Probiotics compared to no probiotics to prevent acute upper respiratory tract infections (URTI) in healthy people**

**Patient or population:** healthy people
**Intervention:** Probiotics
**Comparison:** no probiotics

| Outcomes | Anticipated absolute effects* (95% CI) | | | Relative effect (95% CI) | No of Participants (studies) | Certainty of the evidence (GRADE) |
|---|---|---|---|---|---|---|
| | Risk without probiotics | Risk with probiotics | Difference with probiotics | | | |
| **Number of people who have an URTI** (at least 1 URTI) Follow-up: 3-8 months | **Low**[1] | | | **RR 0.71** (0.47 to 1.07) | 1538 (6 studies) | ⨁⨁◯◯ **low** |
| | **20 per 100** | **14 per 100** (9 to 21) | **6 fewer per 100** (11 fewer to 1 more) | | | |

Note: The authors set a threshold of a small important effect at 5 fewer people with URTI, therefore the effect is a **small important effect**. **Evidence is low certainty** due to some concern with imprecision/inconsistency and risk of bias.

**Options for informative statements**

Probiotics may reduce the number of people with an URTI slightly

The evidence suggests probiotics reduces the number of people with an URTI slightly

Probiotics may result in a slight reduction in URTIs

The evidence suggests probiotics results in a slight reduction in URTIs