How to GRADE the quality of the evidence

The main purpose of this document is to provide instructions for authors on applying GRADE criteria to assess the quality of evidence within reviews.

It contains 3 sections:

1. Instructions to authors
2. Rationale and background material
3. Additional supporting material

This material is intended as a practical supplement to the advice in the Cochrane Handbook. It is intended to assist authors to apply the advice to CCCG reviews in a step-by-step way, and to assist authors with the decisions underpinning GRADE assessments of quality.

We will be seeking ongoing advice from the Cochrane Central Editorial Unit to ensure that this document stays up to date with developing methods.

If you use this resource in preparing your review, please cite it as a reference.

A suggested citation is:

Ryan R, Hill S (2016) How to GRADE the quality of the evidence. Cochrane Consumers and Communication Group, available at http://cccrg.cochrane.org/author-resources. Version 3.0 December 2016.

Acknowledgements:

We gratefully acknowledge the input of Nancy Santesso, Annie Synnot, Claire Glenton, Sue Cole, Miranda Cumpston and Toby Lasserson to this resource, and to Bronwen Merner for assistance in finalising the document.

The advice in this document is adapted directly from advice contained in the Cochrane Handbook (particularly chapter 12), and from resources and advice prepared by the GRADE working group, http://www.gradeworkinggroup.org and https://gradepro.org/ ; see references at the end of this document for further information.

This is a new resource designed specifically for CCCG authors. If you have any comments or suggestions to improve our author resources, please contact Dr Rebecca Ryan, Deputy Coordinating Editor, CCCG at r.ryan@latrobe.edu.au.

Table of Contents

1. Instructions to authors

Before you begin

The GRADE system rates the quality or or certainty[1] of the evidence, and Summary of findings (SoF) tables presents the results (together with the GRADE rating) for the most important outcomes in the review.

The use of GRADE to assess the quality of evidence is **mandatory** for all new reviews. Presentation of the results in a SoF is not mandatory but is strongly encouraged.

Please note that, like any other review method, planning to assess GRADE and the methods needed must be reported as part of the protocol. The methods for creating a SoF table(s) also need to be reported at protocol stage, and you will need to decide which outcomes you will include in the table at the outset.

Please also note that we strongly encourage you to use GRADE to assess the quality of the evidence for **all** outcomes in a review, regardless of whether they are also reported in a SoF table. This is because using GRADE to assess quality helps with writing the results, across outcomes, in a consistent way. This is therefore a very valuable step to prepare for synthesising the results and can help you to develop consistent ways to describe the findings throughout the review.

Please refer to the 'Describing results' document for more on consistent use of language based on assessing both the quality of evidence and size of effects, available at http://cccrg.cochrane.org/author-resources

Finally, please also note that the GRADE ratings of the quality of the evidence should not be confined only to the SoF tables. They GRADE ratings should be reported throughout the review, wherever the findings are presented, as they give a systematic indication of the quality (or certainty) of the evidence on which the findings are based. This will help to ensure that the results are presented consistently throughout the review. The GRADE ratings (and SoF table, if included) can therefore be a valuable tool to help to structure and report the review findings throughout the Abstract, PLS, Effects of interventions and Discussion sections.

[1] Please note that either 'quality' or 'certainty' can be used as terms. There may be less confusion with the use of 'certainty' to describe the outputs of GRADE, as this separates the GRADE assessment from the Risk of bias assessment more clearly, but it is up to you which term you choose to use.

How GRADE works

The GRADE system rates the quality of evidence **for each outcome**, from a rating of HIGH to VERY LOW.

See **section 2.1** for an overview of GRADE and **section 3** for additional sources of information.

GRADE starts with a baseline rating of HIGH for RCTs, and LOW for non-RCTs.

This baseline rating can then be adjusted (downgraded or, less commonly, upgraded) after considering 8 assessment criteria and making a judgement about quality based on these.

Reasons to downgrade the evidence quality

1. Risk of Bias
2. Inconsistency
3. Indirectness
4. Imprecision
5. Publication Bias

For these **5 criteria**, if

- no serious concern exists, do not downgrade quality from the baseline quality (e.g. high for RCTs)
- serious concern exists, downgrade the evidence one level, e.g. from high to moderate (-1)
- very serious concern exists, downgrade the evidence two levels, e.g. from high to low (-2)

Reasons to upgrade the evidence quality (usually not used for evidence from RCTs)

6. Large Magnitude of Effect
7. Dose Response
8. Effect of all plausible confounding factors would be to reduce the effect (where an effect is observed) or suggest a spurious effect (when no effect is observed)

For **criteria 6 to 8**, decide if the evidence should be upgraded once (+1) or twice (+2).

Remember RCT evidence quality is very rarely upgraded.

Decisions to down- or up-grade are not all or nothing, and they rely on your judgement. Sometimes there may be some serious concern but not enough to downgrade by a level of evidence. In such cases, you can decide not to downgrade but should indicate why. In other cases, you may decide that there is some serious concern with 2 or more criteria which adds up to downgrading by a level of evidence.

Based on your assessments, you will decide on a final level of evidence for each outcome, including both meta-analysed and narratively synthesised outcomes. Use this to assign a value for the Quality of evidence.

How to make decisions about what is a concern, or a serious concern, for assessment of each of the 8 criteria, along with detailed instructions about how to work through the GRADE tool are outlined in the sections below.

Remember that readers of your review **must** be able to understand how you reached each quality of evidence rating. You must, therefore, keep track of why you downgraded or upgraded the evidence. You can include the reasons in the results section and, if you plan to report the results of the GRADE assessment in a SoF table, you can explain these decisions in the footnotes to the table. Be transparent about all of the decisions you make. What this means is that for a reader, it should be clear how you reached the judgement about quality of the evidence, based on what you report (in the results section or in a shortened form in footnotes to the SoF table).

The GRADE assessment table (**Section 3 Appendix 1**) can be used to work through a GRADE assessment for each outcome.

You will need one table **for each outcome within each comparison.**

Getting started with GRADE - an overview of the process

1. Decide who will perform the GRADE assessment of quality of the evidence. **Note that at least two review authors** should work independently to assess the quality of evidence and resolve disagreements. The process for reaching consensus where there are disagreements in ratings should be outlined in your Protocol. What we mean by this process is that at least two authors should independently make a decision about each of the GRADE domains, with a process for reaching consensus in place, but that the information on which the decisions are made can be shared, already agreed-upon information that has been extracted for the review. That is, there is no need to independently work out which pieces of information feed into each decision, authors can work from a common starting point, but each making a decision independently about the rating of each GRADE domain.

2. Obtain copies of the table for working through GRADE and become familiar with the decisions that need to be made (see Section 3 Appendix 1). You will need one table **for each outcome**.

3. Select one of the outcomes that represent one of the key outcomes for decision-making irrespective of how many studies contribute data. Predefining the summary of findings table outcomes will help to guard against emphasising results on the basis of the amount of evidence or the size and direction of the effect. It could be synthesised narratively or in meta-analysis (or a combination), or results could come from a single study. Any of these can be assessed with GRADE.

4. For each outcome:
    a. Identify whether the evidence for that outcome comes from:
        i. RCTs (where the rating starts at HIGH quality) or
        ii. non-RCTs (where the rating starts at LOW quality).
    b. Following the guidelines below, systematically work through each of the GRADE criteria, deciding whether to downgrade and/or upgrade the quality of the evidence

and by how much. Upgrading decisions will depend on the nature of the study designs you are basing your assessments on.

c. Keep a comprehensive and transparent record of the reasons for all your decisions about rating the quality of the evidence in the table ('comments' column). The reasons for your decisions about the quality of the evidence form a critical part of the GRADE assessment and must be reported, either as part of the SoF table (footnotes) or in the review if a SoF table is not included.
d. Come to an agreement about the overall quality of the evidence for that outcome.

Assessing the quality of the evidence using GRADE criteria

The GRADE system considers 8 criteria for assessing the quality of evidence.

All decisions to downgrade involve subjective judgements, so a consensus view of the quality of evidence for each outcome is of paramount importance. For this reason downgrading decisions must be made by at least two authors. It requires an assessment of whether any limitations represent a serious threat (in which case it is downgraded by 1 level), or a very serious threat (downgraded by 2 levels) to the validity of the result.

Detailed instructions on how to make judgements about each of these is given below.

Please note that the following advice is adapted from the GRADE Handbook and related materials produced by the GRADE Working Group, available at https://gradepro.org/

Reasons to downgrade the evidence

1. Risk of Bias: what are the limitations?

There are several steps involved in moving from the risk of bias assessment for an individual study to a judgement about the overall risk of bias per outcome. Generally speaking, your assessments of studies using the risk of bias tool forms the basis for identifying limitations at the outcome level. You can then determine how the risk of bias for each of the studies might influence the size, direction, consistency and precision of the overall effect.

This applies both to data synthesised narratively or in a meta-analysis, or results from a single study.

How to GRADE for Risk of Bias

1. Select the outcome to be assessed. Summarising the risk of bias must be done by **outcome**, rather than by study. This is so that the risk of bias for a given result (or outcome), and therefore the confidence we can have in that result, can be determined.

2. Systematically assess the outcome against the following criteria (most are elements of the RCT risk of bias tool) for each of the studies that contribute to it to determine whether the quality of the evidence is affected:
    - Inadequate methods of sequence generation.

- Lack of allocation concealment.
- Lack of blinding of each of:
    - participants,
    - providers,
    - outcome assessors.

  The more subjective an outcome is, the more important effective blinding becomes. For example, symptom improvement is a more subjective outcome than mortality, and is therefore more likely to be biased if unblinded.
- Loss to follow up.

  There is no simple rule of thumb on which to base judgements about this item. The seriousness of losses from a study must be judged based on both the numbers of participants lost **and** the reasons for these losses, looking particularly at whether these are unbalanced across the study groups.
- Failure to follow intention to treat principles in analyses.
- Selective outcome reporting of outcomes and/or analyses.
- Other sources of bias such as stopping the trial for benefit, design specific issues relating to non-standard trial designs, such as cluster or crossover studies.

Note: please also refer to the Risk of Bias section in the CCCG Data Extraction Template (available at: http://cccrg.cochrane.org/author-resources) for more information about making these decisions.

3. Based on the limitations of the studies, come to an overall judgement about whether to downgrade the evidence, and if so, by how much. Table 1 (below) can be used to help make this decision. The following principles may also help to make this judgement.
    - Consider how much each trial contributes to the estimated size of the effect. Usually the larger the trial, with more events, the greater the contribution to the overall effect size.
    - Be conservative when downgrading: you should be fairly confident that most of the information from studies has a substantial risk of bias before you downgrade, bearing in mind that those studies rated as at unclear risk of bias will have a substantial risk of bias associated with them.
    - Make sure you are transparent and clear about why you have reached the decisions that you have, especially if it was a close call.
    - If you have conducted a meta-analysis, consider conducting a sensitivity analysis with only low risk of bias studies. If the effect estimates are unchanged then you may be confident that the risk of bias of the studies does not alter the results.

Note: if the evidence comes from a single RCT, do not automatically downgrade it. You should carefully evaluate the RCT against the GRADE criteria, but it is possible for a very large, well-designed RCT to be rated as high quality. Therefore you need to consider the risk of bias for that one study in making your decision.

Table 1: Moving from risk of bias assessments to overall judgements about limitations

Risk of bias across studies	Considerations	GRADE assessment
Most information is from studies at low risk of bias	Plausible bias unlikely to seriously alter the results	No serious risk of bias, do not downgrade.
Most information is from studies at low or unclear risk of bias.	Plausible bias unlikely to seriously alter the results.	No serious risk of bias, do not downgrade.
	Plausible bias likely to seriously alter the results.	Serious risk of bias, downgrade one level.
The proportion of information from studies at high risk of bias is sufficient to affect the interpretation of results.	Crucial risk of bias for one criterion, or multiple criteria, and likely to seriously alter the results.	Serious risk of bias, downgrade one level.
	Crucial risk of bias for one criterion, or multiple criteria, and likely to very seriously alter the results.	Very serious risk of bias, downgrade two levels.

** Table adapted from Table 12.2.d, The Cochrane Handbook

2. Inconsistency: how consistent are the results?

Heterogeneity refers to any kind of variation across studies, and in systematic reviews different types of heterogeneity can occur:

- Clinical heterogeneity: differences associated with the participants, interventions or outcomes.
- Methodological heterogeneity refers to differences in the way that studies were conducted – for example, differences in study design or risk of bias.

In a review, if studies are judged to be clinically and methodologically similar enough, results may be pooled using meta-analysis, and statistical heterogeneity should then be considered.

Statistical heterogeneity refers to differences in the effects of interventions and arises because of clinical and/or methodological differences between studies. Although some variation in the effects of interventions between studies will always exist, whether this variation is greater than what is expected by chance alone needs to be determined.

Very different estimates of the intervention effect (i.e. heterogeneity or variability in results) across studies suggest that there may be true differences underlying the intervention's effect, that is, that the observed effects are not due only to the effects of the intervention (statistical heterogeneity).

If heterogeneity is present but it has not been possible to identify *why* this variability is present, the quality of evidence should be downgraded by one or two levels, depending on how much inconsistency is present.

Please refer to the CCCG guide on 'Heterogeneity and subgroup analysis' for more on clinical, methodological and statistical heterogeneity, (available at: http://cccrg.cochrane.org/author-resources).

Heterogeneity in results might be adequately explained by sub-group analyses (i.e. by identifying factors that differentially influence the effects of the intervention). Examples might include population characteristics such as age (e.g. older adults might score lower on an outcome than

younger adults), or educational status (e.g. people educated to a higher level might perform better on a test) that might explain some of the variability in the results across studies.

Inconsistency may come about from different sources, such as differences in:

- populations (e.g. sicker populations may respond less well to the intervention)
- interventions (e.g. larger effects with more intense interventions)
- outcomes (e.g. diminishing effects over time).

Looking at how much variability in the results is explained by factors such as these can help to make this decision about how much variability is attributable to identifiable factors, and how much is unexplained.

How to GRADE for inconsistency

1. Consider how much variability there is in the results of studies contributing to the outcome you are assessing.
    - For narrative data, consider whether there is a high degree of inconsistency in the results, such as effects in opposite directions (i.e. benefit and harm), or large variations in the degree to which the outcome is affected (i.e. very large and very small effects).

    - For meta-analysed data, consider whether:
        - there is wide variation in the effect estimates across studies
        - there is little or no overlap of confidence intervals associated with the effect estimates
        - statistical tests that suggest heterogeneity is present, for example:
            - Chi^2 test (testing the null hypothesis that the studies in the meta-analysis have the same underlying effect size) has a low p value
            - I^2 statistic (which quantifies the degree of variability between studies) is large -but please note the I^2 statistic is only one of several things to be considered when assessing heterogeneity, and the thresholds below are only a rough guide.
              As an approximate guide, an I^2 of:
                - 0% to 40% might not be important (low heterogeneity)
                - 30% to 60% might represent moderate heterogeneity
                - 50% to 90% might represent substantial heterogeneity
                - 75% to 100% might represent considerable heterogeneity.
        - whether any heterogeneity has been adequately explained.

2. Decide whether to downgrade on the basis of variability in the results:
    - not at all (inconsistency does not seem to be an issue);
    - one point (some inconsistency exists); or
    - two points (severe inconsistency is present).

See Section 3 for additional resources on rating inconsistency.

3. Indirectness: how do these results apply to my review question?

Indirectness refers to how well the evidence included in the review answers the review question. How applicable is the evidence?

There are two types of indirectness:

1. **Indirect population, intervention, comparator, or outcome**: where the evidence summarised in the review comes from studies that partially address the question of interest to the review, and therefore the conclusions may not be directly answering the review question with respect to:

    - population: might be indirect if included studies were limited to particular participants or settings, for example, if the studies:
        - included only adults, whereas the review looked for studies in all ages (adults and children); or if studies were only found for older adults (65+ years), rather than all adults;
        - included only people with a single disease (not multimorbidities) e.g. people with diabetes alone rather than other chronic diseases or additional conditions.
        - included only people who had completed high school education, rather than people of all educational levels;
        - patients in primary care settings only.

    - intervention: studies only assessed particular versions of the intervention, for example:
        - particular dosing regimens;
        - those delivered by a pharmacist, but not by other professionals;
        - those delivered only once (not multiple times).

    - comparator: studies included comparisons that were not highly applicable, e.g. a comparison group received care that is not currently accepted as usual, standard or routine care.

    - outcome: studies measured outcomes that were not the most informative way of measuring effects of the interventions, e.g.
        - using surrogate outcomes;
        - reporting only endpoint (not intermediate or process) outcomes;
        - reporting outcomes at short but not long term time points.

2. **Indirect comparison**: If a comparison between intervention A and B is not available, a review might compare A with C and B with C. This allows an indirect comparison of the magnitude of effect of A versus B. Such evidence is, however, of lower quality than head-to-head comparisons of A and B would provide.

How to GRADE for indirectness

1. Consider again the question your review set out to address. Did the included studies provide broad answers to the question? Are there restrictions based on what was found, and that might affect applicability of the findings, in terms of:
    - population?
    - intervention?
    - comparator?
    - outcomes?

2. Decide whether the evidence that was found is more restrictive than the review question. If so, then the evidence may not directly answer the review question and you may downgrade for indirectness:
    - not at all (indirectness does not appear to be an issue)
    - one point (some indirectness exists), or
    - two points (indirectness is severe, or there is indirectness from several sources).

When considering the degree of indirectness, bear in mind that these judgements are often not clear cut, and not simply additive. A problem with indirectness of outcomes will often trigger downgrading, but all judgements need careful consideration.

See Section 3 for additional resources on rating indirectness.

4. Imprecision: how precise is the effect size?

Results are imprecise when studies include only relatively few patients or for dichotomous outcomes, there are few events, or when there is a lot of variation in the effects among the participants in continuous measures. As a result, there may be wide confidence intervals (CIs) around the effect estimate.

Note that citing a lack of statistical significance for a result is not a reason for downgrading an outcome for imprecision: instead you should consider the size of the effects, the sample size, the number of events and the CIs around the effect estimate (precision) to make this judgement.

When assessing imprecision, you should look at two things in particular:

1. The number of people analysed: is there enough information to detect a precise estimate of the effect?
2. The CI around the effect estimate: does the CI (i.e. the range of values that the effect estimate might take) include meaningful benefit and harm, or a meaningful effect and no effect (consistent or inconsistent effects)?

How to GRADE for imprecision

1. Assess whether there is enough information (large enough sample size, or large enough number of events) to calculate a precise effect estimate.

    a. For dichotomous outcomes, unless events rates are very low (see also point 2 below) information is likely to be insufficient if:
        - total number of events is less than 300 (a "rule of thumb")
        - total (cumulative) sample size is lower than the calculated optimal information size: i.e. if the total number of participants in the review is less than the number of participants required for a single adequately powered trial

    b. For continuous outcomes information is likely to be insufficient if:
        - total number of participants is less than 400 (a "rule of thumb").

2. Look at the precision of the effect estimate.

    - Do the upper and lower limits include both meaningful benefit and harm (consistent or inconsistent messages) about the effect of the intervention? If the limits of confidence intervals represented the true effect, would they give the same message about the intervention, or not (e.g. does one end indicate a meaningful benefit, and the other no effect or even a harm)?
    - Does the 95% CI (or alternative estimate of precision) around the pooled or best estimate of effect include both little or no effect **and** appreciable benefit or appreciable harm?
        - For dichotomous outcomes, GRADE suggests that the threshold for 'appreciable benefit' or 'appreciable harm' that warrants downgrading is a relative risk reduction (RRR) or relative risk increase (RRI) greater than 25%. Do not overlook absolute effects since for rare events wide confidence intervals of a risk ratio can correspond to small differences in absolute terms.
        - For continuous outcomes, GRADE suggests that the thresholds are the minimal important difference (MID), either for benefit or harm. If the MID is not known, we suggest downgrading if the upper or lower confidence limit crosses the effect size (eg SMD) of 0.5 in either direction.

3. Decide whether there is imprecision in the results, based on your assessments of points 1 and 2 above, and if so, to what extent. Make a decision about whether to downgrade:
    - not at all (imprecision does not appear to be an issue)
    - one point (some imprecision exists), or
    - two points (very serious imprecision exists).

See also Section 3 for additional resources on rating imprecision.

5. Publication Bias: are these all of the relevant studies?

Publication bias is a systematic under or over estimation of the underlying beneficial or harmful effect of the intervention, due to the selective publication of studies or availability of their data. There is good evidence that the nature of the results found in a study determines whether it gets

published. It may be necessary to downgrade for publication bias if the data you are able to include comes from an unrepresentative sample of the studies that have been conducted (e.g. investigators do not report studies because of their results).

One way to assess the likelihood of publication bias is to look at the pattern of study results. This can be visualised using a funnel plot. A funnel plot will display the pattern of results across studies and highlight whether small studies tend to report results in a particular direction compared to larger studies (so-called 'small study effects'). However, there are many reasons why small studies may differ in their findings from larger studies, and so funnel plots should be interpreted carefully.

The risk of publication bias is often higher for reviews that include only small trials, as larger trials are less likely to be unpublished or unknown, irrespective of their actual findings. Please note however that the presence of small studies alone is not necessarily indicative of publication bias: it may reflect a high risk of bias or imprecision in a developing (early) evidence base and is not necessarily confirmation that data has been withheld.

How to GRADE for Publication bias

1. Consider the size of the included studies (and number of events they include). If all results come from small studies, publication bias may be present.

2. Consider constructing a funnel plot, which graphs precision against the size of the effect. If the plot is asymmetrical (skewed) then publication bias may be present. Note, however, that asymmetry of the plot does not always indicate publication bias.

3. As it is difficult to entirely rule out the presence of publication bias, and ways of assessing it are uncertain, the GRADE recommendation is to only downgrade one level at a maximum (not two) on the basis of suspected publication bias. If publication bias is:
    a. undetected, do not downgrade
    b. strongly suspected, downgrade one level.

See Section 3 for additional resources on assessing publication bias.

Reasons to upgrade the evidence

It is **rare** to upgrade the quality of the evidence (e.g. upgrading from a starting rating of low quality for very well-designed and conducted observational studies).

It is **very rare** to upgrade evidence from randomised controlled trials that has been downgraded (i.e. upgrading (based on one of the reasons listed below) a trial that has been downgraded based on risk of bias, imprecision or other criteria).

For observational studies, **only** evidence from studies with no important threats to validity should be upgraded.

There are three major possible reasons to upgrade the quality of evidence.

1. Large magnitude of effect (i.e. strong association)

When methodologically strong observational studies yield large or very large and consistent estimates of the magnitude of an intervention effect, we may be more confident about the results. In situations like this, even though the study design is weak and is likely to overestimate the effects of the intervention, it is unlikely to explain all of the apparent benefit or harm.

2. Dose Response

The presence of a dose-response gradient (relationship) may increase our confidence in the findings of observational studies and thereby increase the quality of evidence.

3. Effect of all plausible confounding factors

On occasion, all plausible confounding from observational studies or randomised trials may be working against the direction of the observed effect, either to:

- reduce the effect seen, or
- increase the effect if no effect was observed.

For example, if only sicker patients receive the intervention, yet they still improve, it may be likely that the actual intervention or exposure effect is larger than the data suggest.

How to upgrade the quality of the evidence:

Consider whether each of the following apply to your outcome:

1. Is there a large magnitude of effect?
    - Large RR >2 or <0.5 (based on consistent evidence from at least 2 studies, with no plausible confounders): upgrade 1 level
    - Very large RR >5 or <0.2 (based on direct evidence with no major threats to validity): upgrade 2 levels
2. Is there a dose-response gradient in the findings?
3. Have all plausible confounding factors been accounted for?
    - Did you find an effect, even though all of the confounding you can think of would have reduced the effect size?
    - Did you fail to find an effect, even though all of the confounding you can think of would have increased the effect size?

Note: only observational studies with no major threats to validity (i.e. that have not been downgraded) can be upgraded.

Refer to Section 3 for additional resources on rating up the quality of evidence.

Incorporating GRADE into the review

Applying GRADE to the assessment of quality of the evidence is a key method in Cochrane reviews. The methods and ratings therefore **must** be adequately described in several sections of the review.

This includes:

1. Describing the **methods** used for assessing the quality of the evidence - not just those for assessing the risk of bias (one component of the quality of the evidence rating).
    - The methods used should be reported under 'Data collection and analysis' in the review.
    - Ideally the methods should be described at protocol stage, along with the selection of outcomes to be presented in Summary of Findings (SoF) tables.
    - Even if the methods were not decided early (i.e. at protocol stage), any methods changed or adopted after the publication of the protocol should be transparently declared in the 'Differences between the protocol and the review' section.
    - The description of methods should include mention of:
        - Use of the GRADE approach (and the GRADE Pro/GDT software if used)
        - How the assessment of quality was rated. For example: We downgraded a starting rating of 'high quality' evidence by one level for serious concerns (or by two levels for very serious concerns) about risk of bias, inconsistency, indirectness, imprecision or publication bias.
        - The number of raters involved.
        - Which outcomes were selected for inclusion in the Summary of findings tables.

    Note: please refer to the CCCG Protocol Template with Guidance (available at http://cccrg.cochrane.org/author-resources) for instructions about how to describe the methods for assessing the quality of evidence using GRADE and preparing SoF tables in your review

2. **Integrating the ratings of quality** with the reporting of results: having used GRADE to assess the quality of the evidence, these ratings need to be reported together with the results.

    See **Section 2.2** for examples.

    - One option is to include a description of the quality rating with the results for an outcome in the 'Effects of interventions' section.

    - Another option is to also include a description of the quality rating together with an explanation of the rating (i.e. major reasons for downgrading) with the results for an outcome in the 'Effects of interventions' section

3. Including information about the quality of the evidence in the **Discussion** section – not just a description about the overall risk of bias.

    See **Section 2.3** for examples.

- This should include an assessment of how the other elements assessed by GRADE (imprecision, inconsistency, indirectness and publication bias), as well as the risk of bias, may affect the confidence you can have in the results.
- This information can be reported in the Discussion under the subheading 'Quality of the evidence'.
- In some cases the reasons for downgrading are similar for many or all of the outcomes and so it may be possible to give an overall description of the quality of the evidence for a set of outcomes.
- In other cases, for example if the downgrading decisions differ markedly between outcomes and the quality of the evidence varies accordingly, a more detailed description of the quality of the evidence may be needed.

4. Brief references to the ratings should also be included in the descriptions of your findings in all summary versions of the results, including the Conclusions, Abstract, Plain Language Summary and Summary of findings tables.

Checklist: final steps

1. **Methods:** have the methods used to rate the quality of the evidence been clearly described in the review (under 'Data collection and analysis)?

   Please refer to the CCCG Protocol Template (available at http://cccrg.cochrane.org/author-resources) for instructions about how to describe the methods for assessing the quality of evidence using GRADE and preparing SoF tables in your review

2. **Explaining the decisions on ratings:** has enough information been provided on the decisions to down or up grade the quality of the evidence?

   This includes reporting:

   a. By **how many levels** the quality of evidence has been down or up graded
   b. **why** the quality of evidence has been down or up graded
   c. This should be described in the 'Effects of interventions' section of the review, for example describing the results together with the quality of the evidence for a given outcome.
   d. How the rating of quality of evidence was arrived at should be transparent so that readers can follow the decisions made.

3. **Using the ratings in the review:** have the ratings of quality of evidence been used throughout the review?

   This assessment of evidence quality is used as the basis for describing results in a consistent and standardised way throughout the review, and to create Summary of Findings (SoF) tables.

   You should use these ratings throughout your review to ensure clear, consistent messages about the effects of the interventions across **all** sections of your review, particularly:

   - Abstract
   - Plain language summary
   - Effects of interventions
   - Discussion
   - Implications for practice, and
   - Summary of findings tables (if you are planning to use these).

More information about how to move from the GRADE ratings of quality of the evidence to consistent wording suitable for inclusion in the review is available at:

- 'Describing results' available at http://cccrg.cochrane.org/author-resources.

2. Rationale and background material

2.1 GRADE – an overview

GRADE is a system for assessing the quality of the evidence of each outcome in a review against eight criteria (including risk of bias, indirectness, inconsistency, imprecision, and publication bias).

In the GRADE system, the quality of evidence for each outcome is graded as HIGH, MODERATE, LOW or VERY LOW.

These assessments reflect the degree of confidence we can have in our effect estimate.

A HIGH rating means that, having assessed all of the potential problems with the evidence quality, we can be confident in our effect estimate. Further research is very unlikely to alter our confidence in that effect estimate.

As we move down the rating list from HIGH to VERY LOW, our confidence in the effect estimate progressively decreases. It becomes increasingly likely that further studies that address some of the problems identified with the quality of evidence will alter the effect estimate.

Note that VERY LOW is the lowest quality rating, no matter how many reasons for downgrading quality there may be.

Table 2: GRADE ratings and their interpretation

Symbol	Quality	Interpretation
⊕⊕⊕⊕	**High**	We are very confident that the true effect lies close to that of the estimate of the effect.
⊕⊕⊕O	**Moderate**	We are moderately confident in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.
⊕⊕OO	**Low**	Our confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect.
⊕OOO	**Very low**	We have very little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.

**Table taken from the GRADE Handbook, available at
http://gdt.guidelinedevelopment.org/app/handbook/handbook.html#h.9rdbelsnu4iy

Authors who are considering an update of their review should contact the editorial base for advice on GRADE and SoF tables if these methods have not been incorporated in previous versions of the review.

2.2 Examples of integrating quality ratings with the Results

Integrating the ratings of quality with the reporting of results: having used GRADE to assess the quality of the evidence, these ratings need to be reported together with the results.

One option is to include a description of the quality rating with the results for each outcome reported in the 'Effects of interventions' section.

For example:

'Skill acquisition: There is moderate quality evidence that multimedia education was more effective than usual care or no education (MD of inhaler technique score 18.32%, 95% CI 11.92 to 24.73, two studies with 94 participants) and written education (risk ratio (RR) of improved inhaler technique 2.14, 95% CI 1.33 to 3.44, two studies with 164 participants). There is very low quality evidence that multimedia education was equally effective as education by a health professional (MD of inhaler technique score -1.01%, 95% CI -15.75 to 13.72, three studies with 130 participants).'

Ciciriello et al. Multimedia educational interventions for consumers about prescribed and over-the-counter medications. Cochrane Database of Systematic Reviews 2013, Issue 4. Art. No.: CD008416. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008416.pub2/pdf

Another option is to also include a description of the quality rating together with an explanation of the rating (i.e. major reasons for downgrading) with the results for an outcome in the 'Effects of interventions' section.

2.3 Examples of using quality ratings in the Discussion

Including information about the quality of the evidence in the **Discussion** section – not just a description about the overall risk of bias.

- This should include an assessment of how the other elements assessed by GRADE (imprecision, inconsistency, indirectness and publication bias), as well as the risk of bias, may affect the confidence you can have in the results.
- This information can be reported in the Discussion under the subheading 'Quality of the evidence'.
- In some cases it may be possible to give an overall description of the quality of the evidence for particular outcomes.

For example:

'Using GRADE, we assessed the certainty of the evidence to be moderate to low for outcomes for which data were available. The reasons for these judgements are outlined in the Summary of findings for the main comparison. We assessed the two included trials as being at low risk of bias. However, there were sparse data for several outcomes indicating that further trials are needed.'

Saeterdal I, Lewin S, Austvoll-Dahlgren A, Glenton C, Munabi-Babigumira S. Interventions aimed at communities to inform and/or educate about early childhood vaccination. *Cochrane Database of Systematic Reviews* 2014, Issue 11. Art.No.: CD010232.

'**Quality of the evidence**
…we found the quality of evidence for most outcomes to be of moderate or low quality primarily due to risk of bias and imprecise results because of few fracture events. It could be argued that evidence for hip fractures in the community, and for pelvic fractures, could be assessed as higher quality since the incidence of events is very low and the confidence intervals narrow enough that additional research would not be required. However, the unexplained heterogeneity across studies for pelvic fractures (in particular due to the O'Halloran 2004 study) warrants additional research to determine the effects of hip protectors on pelvic fractures and the evidence was therefore assessed as low quality.'

Santesso N, Carrasco-Labra A, Brignardello-Petersen R. Hip protectors for preventing hip fractures in older people. Cochrane Database of Systematic Reviews 2014, Issue 3. Art. No.: CD001255.

In other cases, for example if the quality of the evidence is very variable across outcomes, a more detailed description of the quality of the evidence may be needed.

For example:

'Conclusions regarding the superior effect of multimedia education on knowledge were based on the results from six studies containing 817 participants. The evidence was downgraded to low quality due to the studies having an unclear risk of bias for allocation concealment, blinding of outcome assessors or both, and due to considerable statistical heterogeneity (I2 = 89%).'

'Conclusions regarding the superior effect of the addition of multimedia education to a co-intervention on knowledge was based on the results of 2 studies with 381 participants. The evidence was of moderate quality and was downgraded due to the studies having unclear risk of bias for allocation concealment. The lack of effect on skill acquisition of the addition of multimedia education to written education was based on a single study of 87 participants. The evidence was of very low quality due to the single study having unclear risk of bias for allocation concealment and due to a wide 95%CI that included both no effect and substantial effect in the direction of the multimedia group.'

Ciciriello et al. Multimedia educational interventions for consumers about prescribed and over-the-counter medications. Cochrane Database of Systematic Reviews 2013, Issue 4. Art. No.: CD008416. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD008416.pub2/pdf

3. Additional supporting material

General information

More information about GRADE and Summary of findings (SoF) tables is available at:

- Chapters 11 and 12 of the Cochrane Handbook
- The GRADE Handbook, available in GRADEPro software (used to develop the SoF table), available at http://www.gradepro.org (Under 'Learn GRADE Methodology')
- Online at http://www.gradeworkinggroup.org/publications/JCE_series.htm

Please also refer to the following CCCG documents for specific guidance, available at http://cccrg.cochrane.org/author-resources:

- 'Preparing Summary of Findings (SoF) tables
- 'Describing results'

Assessing risk of bias

- CCCG Review Template with Guidance contains explanation about how to assess each of the items of the risk of bias tool; available at http://cccrg.cochrane.org/author-resources
- CCCG Data extraction Template contains detailed decision rules to assist you with assessing each item of the risk of bias tool; available at http://cccrg.cochrane.org/author-resources
- The Cochrane Training website has an interactive training module on risk of bias at http://training.cochrane.org/resource/risk-bias-online-learning-module, and an online presentation summarising Cochrane guidance at http://training.cochrane.org/resource/assessing-risk-bias-included-studies.

Rating inconsistency

- Guyatt GH, Oxman AD, Kunz R, Woodcocke J, Brozeka J, Helfand M, et al (2011). GRADE guidelines:7. Rating the quality of evidence – inconsistency. J ClinEpidemiol.2011 Aug 2, available at http://www.gradeworkinggroup.org/publications/JCE_series.htm
- The Cochrane Handbook Chapter 9, Section 9.5
- Heterogeneity and subgroup analyses in Cochrane Consumers and Communication Review Group reviews: planning the analysis at protocol stage, available at http://cccrg.cochrane.org.

Rating indirectness

- Guyatt GH, Oxman AD, Kunz R, Woodcocke J, Brozeka J, Helfand M, et al (2011). GRADE guidelines:8. Rating the quality of evidence – indirectness. J Clin Epidemiol. 2011 Aug 1, available at http://www.gradeworkinggroup.org/publications/JCE_series.htm
- Cochrane Handbook Chapter 12, Section 12.3

Rating imprecision

- Guyatt GH, Oxman AD, Kunz R, Brozeka J, Alonso-Coello P, Rind D, et al (2011). GRADE guidelines:6. Rating the quality of evidence - imprecision. J Clin Epidemiol. 2011 Aug 12, available at http://www.gradeworkinggroup.org/publications/JCE_series.htm
- Cochrane Handbook Chapter 12, Section 12.4.

Assessing publication bias

- Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozeka J et al (2011) GRADE guidelines:5. Rating the quality of evidence –publication bias. J Clin Epidemiol. 2011 Aug 1, available at http://www.gradeworkinggroup.org/publications/JCE_series.htm
- The Cochrane Handbook Chapter 10.

Rating up the quality of evidence

- Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al (2011). GRADE guidelines:9. Rating up the quality of evidence. J Clin Epidemiol. 2011 Aug 1, available at http://www.gradeworkinggroup.org/publications/JCE_series.htm

Describing the methods for using GRADE criteria and/or preparing summary of findings tables in CCCG reviews

- Refer to the CCCG Protocol template with Guidance, available at http://cccrg.cochrane.org/author-resources.

Appendix 1: Table for assessing the GRADE criteria

This table can be used to work through a GRADE assessment for each outcome i.e. you will use one such table per outcome (within each comparison).

GRADE criteria	**Rating** (circle one)	**Footnotes** (explain reasons for down- or upgrading)	**Quality of the evidence** (Circle one)
Outcome:			
Study design	RCT (starts as high quality) Non-RCT (starts as low quality)		
Risk of Bias *(use the Cochrane Risk of Bias tables and figures)*	No serious (-1) very serious (-2)		⊕⊕⊕⊕ High
Inconsistency	No serious (-1) very serious (-2)		⊕⊕⊕◯ Moderate
Indirectness	No serious (-1) very serious (-2)		⊕⊕◯◯ Low
Imprecision	No serious (-1) very serious (-2)		⊕◯◯◯ Very Low
Publication Bias	Undetected Strongly suspected (-1)		
Other (upgrading factors, circle all that apply)	Large effect (+1 or +2) Dose response (+1 or +2) No Plausible confounding (+1 or +2)		