Australian Government
National Health and Medical Research Council

NHMRC

August 2022

Assessing certainty of evidence using GRADE

Prepared by NHMRC for the Department of Health and Aged Care

Purpose

The purpose of this document is to provide the Department of Health and Aged Care (the Department) with:

- an overview of GRADE[1]
- a guide to interpreting how GRADE was applied to assess the certainty of evidence for the Review of Natural Therapies[2].

Background

What has the Department asked NHMRC to do?

NHMRC has been engaged by the Department to evaluate the evidence for the clinical effectiveness of 16 natural therapies excluded from private health insurance rebates on 1 April 2019 via a series of 16 evidence evaluation reports.

NHMRC has commissioned independent evidence reviewers to conduct the evidence evaluations and has appointed the Natural Therapies Working Committee (the Committee) to oversee the evaluations.

What has NHMRC asked the independent evidence reviewers to do?

Independent evidence reviewers have been commissioned to align each evidence evaluation report with the methodology outlined in the *Cochrane Handbook for Systematic Reviews of Interventions* (Cochrane handbook),[3] where applicable and pragmatic.

The Cochrane Handbook recommends that GRADE[1] be adopted to assess the certainty (or quality/ strength) of an evidence base as part of a systematic review. GRADE is also recommended by NHMRC for development of evidence-based products, such as guidelines and was recently adopted by the *Australian Technical Advisory Group on Immunisation* (ATAGI) in development of the Australian Immunisation Handbook[4]. GRADE has been utilised by the independent evidence reviewers to assess the evidence base for each of the 16 natural therapies.

For the purposes of this document, the term 'evidence evaluation report' is interchangeable with the term 'systematic review' (as defined in the Cochrane Handbook, Chapter 1: *Starting a review*)[3].

---

[1] GRADE: Grading of Recommendations, Assessment, Development and Evaluation. Detailed information about GRADE is available at www.gradeworkinggroup.org

[2] Department of Health and Aged Care, Natural Therapies Review 2019-20. Accessible at: https://www.health.gov.au/health-topics/private-health-insurance/private-health-insurance-reforms/natural-therapies-review-2019-20

[3] Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook

[4] Development of the Immunisation Handbook https://immunisationhandbook.health.gov.au/contents/about-the-handbook/development-of-the-handbook

What is GRADE and why use it to assess the certainty of evidence?

GRADE is an internationally recognised framework and tool used for:

1. grading the certainty (quality or strength) of evidence in systematic reviews
2. providing a systematic approach for making decisions or recommendations about treatments (or interventions) using the best available evidence.

Figure 1 (below) outlines the overall GRADE process from developing a research question, rating the certainty of the evidence, through to making a recommendation. This document relates to the 'systematic review' section in **Figure 1** (Evidence synthesis and Rate certainty of evidence).

The remaining areas outlined in **Figure 1** (Balance of consequences and

Recommendations) relate to the development of recommendations or decisions under a GRADE Evidence to Decision framework. This is described in more detail in the Evidence to Decision document provided to the Department in August 2022.

Figure 1: *Overview of the GRADE process*

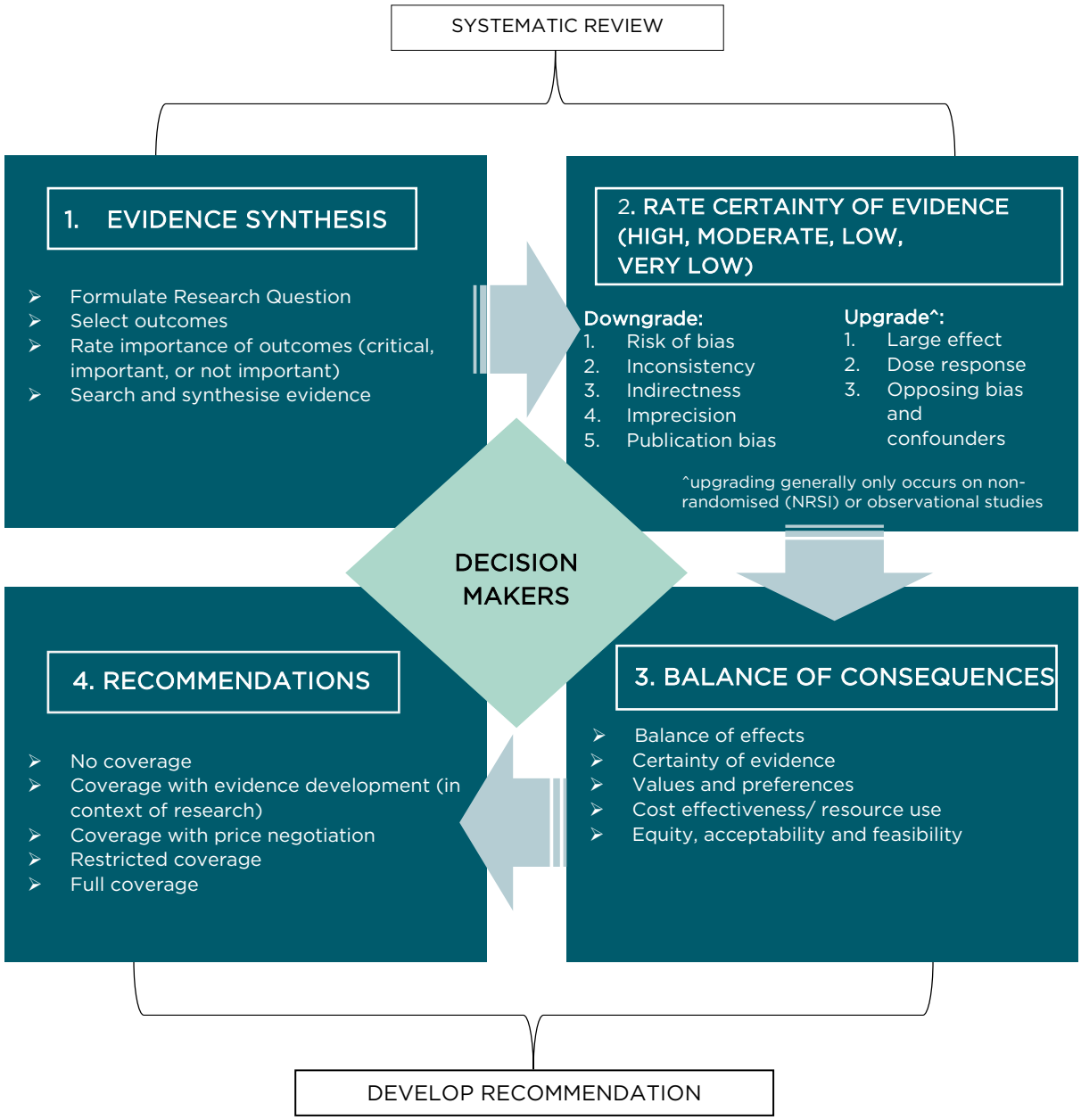

Image adapted from: Muhammad, Rafiq & Boccia, Stefania (2018) available at: https://pubmed.ncbi.nlm.nih.gov/29410601/

The GRADE process aims to improve transparency and consistency in reporting and decision making by assessing key aspects of the way studies are designed, run and analysed, which affect how certain (or confident) a reviewer can be that the results reported in studies are accurate. The GRADE process for rating certainty of evidence, is designed to be repeatable and transparent. It also formalises which aspects of the methods and results of studies to critically appraise.

Using the GRADE process, a rating of certainty is given for each pre-specified critical or important outcome in a systematic review, describing it as high, moderate, low or very low certainty (refer to **Table 1** below).

Table 1: *Grade ratings of certainty*

High certainty ⨁⨁⨁⨁	*The authors have a lot of confidence that the true effect is like the estimated effect*
Moderate certainty ⨁⨁⨁	*The true effect is probably close to the estimated effect*
Low certainty ⨁⨁	*The true effect might be markedly different from the estimated effect*
Very low certainty ⨁	*The true effect is probably markedly different from the estimated effect*

Estimated effect: how much a treatment (or intervention) impacts an outcome (e.g. pain) when compared to people who do not receive the treatment (or intervention) in a study.

True effect: represents the effect a treatment (or intervention) is likely to have on the general population (outside of a study).

Assessing certainty of evidence using GRADE[5]

To begin the GRADE process for assessing the evidence for certainty (quality or strength), a panel defines a research question and selects the population, intervention (or treatment), comparison and outcome/s of interest, this is termed a PICO.

Defining a PICO is essential in determining (1) what to search for when looking for evidence and (2) what data to select (or extract) from studies that meet the PICO criteria[6].

[5] NHMRC's Guidelines for Guidelines website includes an overview of 'assessing certainty of evidence' for guideline developers, accessible at: https://www.nhmrc.gov.au/guidelinesforguidelines/develop/assessing-certainty-evidence.

[6] For the Natural Therapies review, the Department and its Natural Therapies Expert Advisory Panel (NTREAP) were asked to pre-specify the PICO for each of the 16 natural therapy reviews. Given a PICO was unable to be pre-specified, all populations and outcomes were stated as important. To ensure the reviews were manageable and achievable within a reasonable timeframe and budget, NHMRC and its Natural Therapies Working Committee (the Committee) developed processes for population prioritisation (where required) and outcome prioritisation, consulting NTREAP on each. Whilst the processes made the reviews more manageable, standard Cochrane Systematic Reviews usually target a review to a small number (<10) of related populations (e.g. Pilates for low back pain). In contrast some of the natural therapy reviews have included >20 (unrelated) populations of interest, making the review process more complex than a standard Cochrane Systematic Review.

Under GRADE, the certainty of the evidence is assessed across outcome/s of interest (the 'O' in PICO). The results data from outcomes considered 'critical' or 'important' by a panel are then extracted from eligible studies (if available), combined across studies in GRADE summary of findings tables and assessed for certainty. When multiple studies assess the same critical or important outcome their data is combined to produce an overall estimate of effect (using a meta-analysis). Outcomes reported in studies that are considered low importance by a panel are not included in the systematic review or summary of findings tables.

To assess the certainty of the evidence for critical or important outcomes, GRADE has identified **five key domains** that are considered in totality for each outcome, including:

- Risk of Bias
- Inconsistency
- Indirectness
- Imprecision
- Publication bias

As each domain is assessed, the level of certainty for an outcome can be downgraded. When assessing the effects of an intervention, randomised trials begin with *high certainty* as a default rating. Randomised trials are considered best practice for assessing effectiveness, as the randomisation process reduces bias when examining cause-effect relationships between interventions and outcomes.

In contrast, non-randomised trials and observational studies begin with a *low certainty* default rating, unless being assessed for risk of bias using the ROBINS-I tool (which starts non-randomised and observational studies at high certainty). The ROBINS-I tool was developed with the recognition that non-randomised and observational studies are often the main source of evidence available to assess many public health interventions, where it may be inappropriate (or impractical) to conduct a randomised controlled trial, but where it is still important to assess the certainty of studies based on their merits.

Upgrading generally only occurs for non-randomised studies and observational studies. For non-randomised and observational studies upgrading can be considered using three additional domains:

- Large effects
- Dose-response
- Opposing bias and confounders

Domains that may <u>decrease</u> certainty in the evidence

1. Risk of Bias

Assessing limitations of the study design and reporting

<u>Key features</u>

- Risk of Bias (RoB) is used to assess whether the design features and/or conduct of a study have led to misleading or 'biased' results.
- Bias could either be in favour of an intervention or in favour of a control/comparison.
- Study authors may or may not be aware of bias when developing a study.
- GRADE uses a two staged process to assess risk of bias
  1. Assessing risk of bias for individual studies (using specific RoB tools)
  2. Assessing the overall RoB across multiple studies for critical and important outcomes.

Assessing overall risk of bias across multiple studies

- For Cochrane systematic reviews, the RoB2 tool is recommended to assess randomised controlled trials (RCT) and ROBINS-I tool is used for non-randomised studies of interventions (NRSI).
- When a meta-analysis is possible, each individual risk of bias result is combined to produce an overall risk of bias for each critical or important outcome.
- A review author would generally place more emphasis on the risk of bias result of the study/s that contributed the most weight to the meta-analysis.

GRADE ratings for overall RoB

- Low RoB – no downgrading of certainty occurs
- Unclear RoB (but seems likely) – downgrading of certainty may occur
- Some concerns of RoB – downgrading of certainty is likely to occur
- High RoB– downgrading of certainty occurs by one or two levels.

2. Inconsistency

Assessing whether the results from different studies are similar or different

Key features

- Inconsistency (or heterogeneity) assesses whether the results of studies are consistent with each other.
- When the reported estimate of effect (either for or against an intervention) is consistent across multiple studies, a systematic review author can be more certain (or confident) that the reported effect is likely to be true.
- In contrast, when reported estimates of effects are different across studies, systematic review authors are less certain (or confident) that the reported effect is true or accurate.

Options to assess results of studies for inconsistency

- Using visual cues such as confidence interval on a forest plot – a forest plot is a visual display of the results for each included study, as well as the combined results for all included studies (i.e. results of the meta-analysis). The result (estimate of effect) of each study is depicted as a square and the confidence intervals (depicted as a line) show how much difference there is within each study. The estimate of effect across studies is depicted as a diamond.
- Using statistical tools to calculate inconsistency – the statistical tool I^2 can be used to assess the percentage of variation across studies that is due to inconsistency, rather than chance. The p value of the test for heterogeneity is then measured to check if the inconsistency is statistically significant or not.
- Conduct subgroup analysis (i.e. test whether an intervention affects different groups of people differently e.g. young vs older people) to investigate possible sources of high heterogeneity. The groups for these tests should be specified at the protocol stage.

GRADE ratings for inconsistency

- Where there is only one study included for an outcome, downgrading for inconsistency does not occur.
- Downgrading can occur if the results across studies (in a meta-analysis):
    - Have wide variance of effect estimates
    - Show minimal (or no) overlap of confidence intervals
    - Show heterogeneity/ inconsistency as indicated by statistical measures.

3. Indirectness

Assessing whether the study results can be applied to the population of interest

Key features

- A well-designed study seeks to include a mix of people that mirror the general population of interest. This is to ensure that the inferences drawn from the results are likely to be applicable to that population (i.e. direct).
- In contrast, if a study only includes people that do not mirror the general population of interest, the inferences drawn from results are less likely to be applicable or relevant (i.e. indirect) to the population they are designed to impact.

Assessment of indirectness in a systematic review

- There are four criteria to assess how direct the evidence in a systematic review is to the pre-specified PICO criteria, including:
  1. Differences in study populations - how applicable the included participants and settings are to the overall research question and population of interest
  2. Differences in interventions or differences in delivery of interventions
  3. Differences in outcome measures (surrogate outcomes)
  4. Indirect comparisons

GRADE ratings for indirectness

- For the Review of Natural Therapies, downgrading for differences in population is generally not required given that most of the therapies have populations prioritised to be applicable to the Australian context, and in addition many of these therapies are likely to have a broad mechanisms of action across populations.
- Downgrading may be applied if the intervention (therapy) was not delivered in the usual way (e.g. one session of Rolfing when the Rolfing method involves a set of ten specific sessions).

4. Imprecision

Assessing how much uncertainty there is within studies, and how big effects are

Key features

- The precision (or imprecision) of the results of a study is related to the number of participants (i.e. sample size) and events and is assessed by the confidence interval around the combined estimate of effect.
- When assessing imprecision, it is important to consider where the upper and lower limit of the confidence interval sits in relation to the threshold (i.e. what is the minimum difference required) for interpreting an outcome. In a systematic reviews imprecision is rated per outcome.

Assessment of imprecision in a systematic review

- The GRADE approach rates imprecision using the confidence interval for the overall combined meta-analysis result (the width of the diamond on a forest plot). If the diamond is wide on the forest plot, this generally indicates imprecision. What is defined as “wide” depends on the scale of the forest plot and the thresholds being used. For example, a width of 2 is wide on a scale of 0-5, but not on a scale of 0-100.

- Statistical methods are also recommended to assess imprecision (e.g. calculation of the optimal information size, or rule of thumb of >400 participants).

GRADE ratings for imprecision

- When downgrading for imprecision, consideration is given to:
    - whether the combined confidence interval crosses the threshold for comparison – i.e. either no difference between the intervention and control or a pre-defined threshold (or minimal clinically important difference MCID) that looks at whether the effect is large enough to be useful to people who want to use the intervention.
    - Outcome variables with only one study, and/or less than the optimum information size. For continuous outcomes (most of the outcomes in the Natural Therapies Review) Cochrane recommends a sample size >400 to be confident the optimum information size is met. This is because the results of small studies can be different just by chance; that is, if you run the same small study multiple times you may get differing results.

5. Publication bias

Assessing whether some results are probably not published

Key features

- Publication bias refers to the selective publication of studies that may over-estimate the effect of an intervention.
- Publication bias is generally more likely to occur for research showing an effect (and for outcomes with a desirable effect) versus research showing nil or minimal effect (and for outcomes that show an undesirable effect). Similarly, publication bias is more likely for studies with small sample sizes and when the people running the studies have a commercial interest in the outcome of a study.

GRADE assessment of publication bias in a systematic review

- Statistical tests (e.g. funnel plot) can be conducted by systematic reviewers to assess publication bias, but only when 5-10 or more studies are available to combine in a meta-analysis.
- A comprehensive search is important to assess publication bias. GRADE recommends suspecting publication bias if the only published studies are small studies with positive results.
- Sometimes it is possible to find results (and usable data) which have not been formally published (e.g. grey literature). However grey literature may not be peer-reviewed for accuracy and completeness so data should be used with caution.
- Clinical trial registries contain trials that have been registered
    - some trials may have been completed and published, some may have been started, and some may have stopped due to lack of funding or interest.
    - Researchers do not always update trial registries, so it is not always possible to know which studies are truly ongoing and which studies have stopped.
    - If there are a lot of studies which are listed as complete, but which have not been published, this may be evidence of publication bias.


NHMRC

Domains that may increase certainty in the evidence

To be considered for upgrading, non-randomised and observational studies must be rated low for risk of bias on domains relating to missing data or selective reporting, with no other concerns on other GRADE domains (i.e. downgrading for RoB would only occur by one level due to randomisation, and not for other RoB signalling questions, or for other GRADE domains).

1. Large effects

GRADE advises that upgrading for large effects is rarely conducted and if so, should be done with caution and a sound rationale.

Assessing whether there is a large magnitude of effect

- The GRADE working group suggest that a relative risk greater than 2 can be upgraded by 1 level and a relative risk greater than 5 can be upgraded by 2 levels. Relative risk is a way of measuring how likely something is in one group versus another. A relative risk of 2 means that people in the intervention (or treatment) group are twice as likely as people in the control group to show the outcome effect. Relative risk is best applied to outcomes which are dichotomous (e.g., pain versus no pain) rather than continuous (e.g., rating pain on a scale of 1-10).
- There is no guidance for other measures of effect about what would be a large enough effect to justify upgrading.
- Some other effect size measures have general guidelines about what is considered a "small", "medium" or "large" effect but these were not developed in consideration with GRADE. Statisticians caution that such generic guidelines about effect sizes should only be used in new areas where there is no information about what a clinically meaningful difference would be.

2. Dose-response gradient

For the Review of Natural Therapies, the evidence evaluations are not comparing different amounts of any of the interventions, so upgrading NRSI or observational studies for a dose-response gradient is not applicable.

Assessing whether there is a dose-response gradient

- The dose-response gradient (or dose dependant response) refers to the effects of a substance (such as a drug) on the treatment of an outcome (e.g. reduction in pain).
- A dose-dependent gradient is seen when the effect of a substance changes when the dose of the substance is altered (either increased or reduced). For example, for substances such as essential nutrients (e.g. potassium), the dose response curve is U-shaped, where not enough potassium impacts health outcomes, as does having too much.
- Where a dose-response gradient is seen, it is assumed that there is a strong cause-effect relationship. This cause-effect relationship can increase the confidence in the observed results of observational and NRSI studies.
- Where there are no additional concerns with other GRADE assessment criteria, upgrading for a dose-response gradient is considered appropriate.

3. Opposing bias and confounders

The final case when upgrading of observational studies may be appropriate is when there may be differences between groups in a study before the study starts, which lead to the results being hidden.

- In RCTs, randomisation should remove differences between groups before the intervention. In observational studies randomisation does not occur and it may not be possible to match groups before the study starts.
- For example, a study comparing mortality rates in private and public hospitals may start out with patients being treated for different conditions at the two kinds of hospitals, so comparing mortality rates may be problematic.
- If it can be determined that the starting differences between groups are likely to be hiding an effect, then upgrading may be appropriate.

Conclusion

This document provides the Department with an overview of the GRADE approach to assessing the certainty of an evidence base. The document is intended to aid in interpreting how GRADE was applied to assess the certainty of evidence for the Review of Natural Therapies.

In summary, GRADE can be applied to assess the certainty of evidence across outcomes in systematic reviews. GRADE is a transparent and repeatable tool to provide more confidence in the results of studies included in systematic reviews.

References and resources

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.3 (updated February 2022). Cochrane, 2022. Available from: www.training.cochrane.org/handbook

Schünemann H, Brożek J, Guyatt G, Oxman A, editors. GRADE handbook for grading quality of evidence and strength of recommendations. Updated October 2013. The GRADE Working Group, 2013. Available from: guidelinedevelopment.org/handbook

NHMRC's Guidelines for Guidelines website includes an overview of 'assessing certainty of evidence' for guideline developers, accessible at: https://www.nhmrc.gov.au/guidelinesforguidelines/develop/assessing-certainty-evidence.