Rating Certainty of the Evidence Using GRADE Guidelines

Abhijeet Rakshasbhuvankar

Abstract Systematic reviews in healthcare should review and synthesise all available evidence, and provide information regarding certainty (quality) of evidence to inform readers about the amount of confidence they can place in the evidence. Many international organisations, such as the World Health Organisation, National Institute for Health and Care Excellence (NICE) and the Cochrane Collaboration have recommended GRADE (The Grading of Recommendations Assessment, Development, and Evaluation) guidelines to rate the certainty of (a body of) evidence in systematic reviews. These guidelines provide a structured and transparent process to rate the certainty of evidence considering critical factors which may decrease (risk of bias, inconsistency, indirectness, imprecision, and reporting bias) or increase (a very large effect, dose-response relation, and bias that would decrease effect estimate) our confidence in effect estimates. The process of rating certainty of the evidence is presented as a Summary of Findings table in a systematic review. This chapter covers the use of GRADE guidelines for rating certainty of evidence in a systematic review.

Keywords Certainty · Evidence · GRADE · Imprecision · Inconsistency · Indirectness · Publication bias · Quality · Reporting bias

Introduction

Systematic reviews aim to synthesise the available evidence to help clinicians, guideline developers, and researchers make evidence-based decisions, develop clinical care guidelines, and identify the gaps in knowledge, respectively. The synthesis should include not only the effect estimates but also the level of confidence in them. The level of confidence in the effect estimate decides its usefulness and is determined by the certainty (quality) of evidence (Guyatt et al. 2008). The

A. Rakshasbhuvankar (✉)
School of Medicine, University of Western Australia, Perth, WA 6008, Australia
e-mail: Abhijeet.rakshasbhuvankar@health.wa.gov.au

S. Patole (ed.), *Principles and Practice of Systematic Reviews and Meta-Analysis*,
https://doi.org/10.1007/978-3-030-71921-0_10

certainty of the evidence is defined as the extent to which one can be confident that an estimate of effect is correct (Atkins et al. 2004). Various systems have been used to grade the certainty of evidence.

The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) guidelines were developed by the GRADE Working Group and are recommended by the Cochrane collaboration to rate the certainty of evidence (Puhan et al. 2014; Schunemann et al. 2019). The GRADE system has an advantage over other systems. It explicitly considers multiple vital components that determine the evidence's certainty, provides a structured and explicit approach for reviews to make their judgments, and enables readers to understand the reasoning behind the decisions.

In addition to rating the certainty of the evidence, GRADE guidelines are also used for rating strength of recommendations. The strength of recommendation indicates the extent to which one can be confident that adherence to the recommendation will do more good than harm (Atkins et al. 2004). The judgment regarding the strength of recommendation in addition to the certainty of the evidence requires careful consideration of the balance between beneficial versus harmful effects, baseline risk, and available resources. This chapter covers the approach to the judgment about the certainty of evidence using the GRADE system.

GRADE Levels of Evidence

GRADE classifies certainty of evidence in four levels: high, moderate, low, and very low (Puhan et al. 2014). The level of confidence progressively decreases as we move stepwise from "high" to "very low" category (Fig. 1) (Guyatt et al. 2008). In general, the certainty of evidence generated from randomised controlled trials (RCTs) is considered as "High," and that from observational studies is regarded as "low." However, significant concerns regarding any of the following factors may downgrade certainty of evidence: risk of bias (ROB), inconsistency, indirectness, imprecision, and publication bias. The certainty of evidence may be upgraded, although rarely, in observational studies in the presence of large effect size, dose-response gradient, or plausible confounders or biases that increase the confidence in the estimated effect (Guyatt et al. 2008; Balshem et al. 2011). The details regarding the factors which can downgrade or upgrade certainty of evidence in a systematic review are described below.

Risk of Bias (ROB)

Bias is a systematic error in results and arises from methodological flaws in a study (Higgins et al. 2019). The reliability of RCT results depends on the extent to which potential sources of biases have been avoided. Bias may arise from the

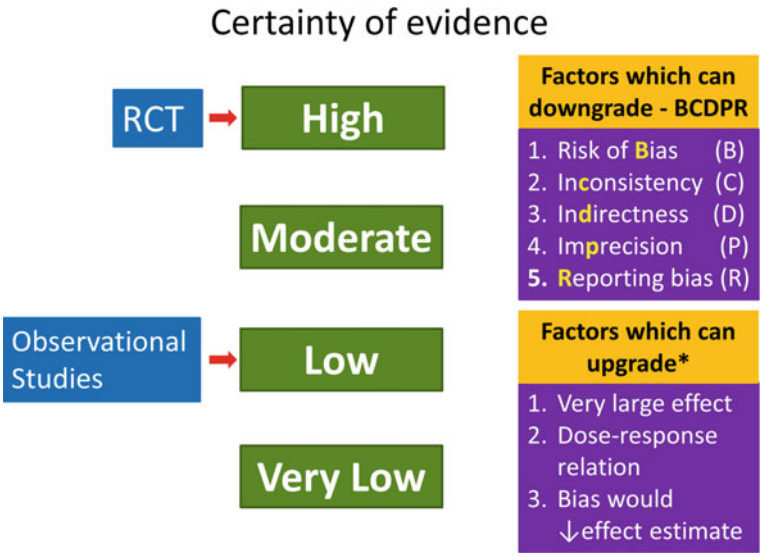

Fig. 1 Levels of certainty of evidence and the factors which downgrade and upgrade it

randomisation process, deviations from intended interventions, missing outcome data, measurement of the outcome, and selection of the reported result. Risk of bias (ROB) assessment is an integral part of the systematic review methodology. GRADE requires the systematic reviewers to decide the (overall) ROB for each outcome across all studies and all domains. The judgment demands careful consideration of ROB in the individual studies for the outcome under consideration and the extent to which the study contributes to the effect estimate (weightage).

(1) *ROB assessment*: Each outcome under consideration is assessed for five sources (domains) of ROB. The risk in each domain is judged as Low risk, Some concerns, or High risk. The details of evaluating an individual study for the ROB are provided elsewhere in this book.
(2) *Contribution (weightage) of the study to the effect estimate*: The contribution of a study for the ROB in a systematic review is proportional to the contribution the study makes for the effect estimate. For example, Fig. 2 shows a forest plot and ROB for a hypothetical systematic review of drug A for pancreatic cancer for the outcome of five-year survival. Studies A, C, and F have high ROB from multiple sources; however, the forest plot indicates that the studies contribute to a negligible extent to the pooled effect estimate. In contrast, Studies B and E, which add the most to the pooled estimate, have low ROB. Hence the reviewers may judge ROB for drug A in pancreatic cancer for the outcome of five-year survival as "Low".

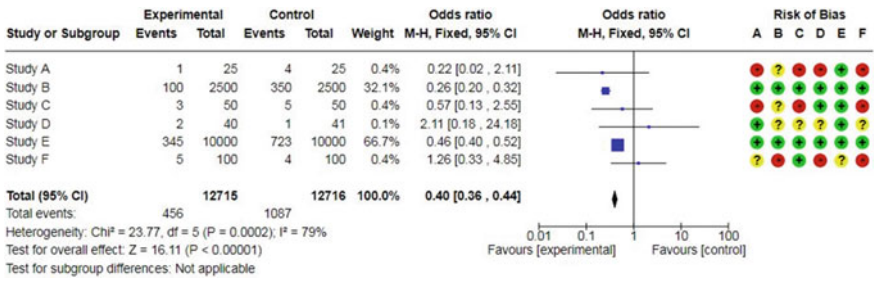

Fig. 2 Hypothetical forest plot and risk of bias—judgement regarding overall risk of bias

Suggestions for downgrading for ROB: (1) If most information is from studies at low ROB: Do not downgrade; (2) If most information is from studies with some concerns: Downgrade by one level, (3) If most information is from studies at high ROB: Downgrade by one or two levels based on the seriousness of limitations.

Reviewers need to apply judgement while deciding overall ROB. In close-call situations, reviewers should be conservative in the decisions of rating down the evidence, should consider ROB judgement in the context of other limitations, and make explicit statements regarding the reasoning behind their judgement (Guyatt et al. 2011).

Inconsistency (Heterogeneity)

Consistency in a systematic review refers to the similarity in the magnitude of effect estimates of the studies. The study results are inconsistent when the variations in the effect estimates between the studies cannot be explained based on chance alone. Inconsistency which cannot be explained by a priori hypotheses may decrease our confidence in the results. Inconsistency is important only when it reduces our confidence in the effect estimates. Assessment of inconsistency of effects across the studies is an integral part of a meta-analysis and grading of evidence (Higgins et al. 2003).

Judgement regarding inconsistency is based on visual inspection of forest plot and statistical tests (Guyatt et al. 2011).

(1) *Forest plot*. The direction of effect and overlap of confidence intervals between the trials are two critical factors in the forest plot, which help in the judgement regarding inconsistency. The impact of these two factors on the judgment of inconsistency is explained with the help of hypothetical forest plots in Fig. 3. In the forest plot A, the directions of effects in the first two studies are different from those in the second two studies. However, the magnitude of the difference is small, and the confidence intervals of the trials overlap. Therefore, the forest plot does not show inconsistency, and our confidence in the pooled effect

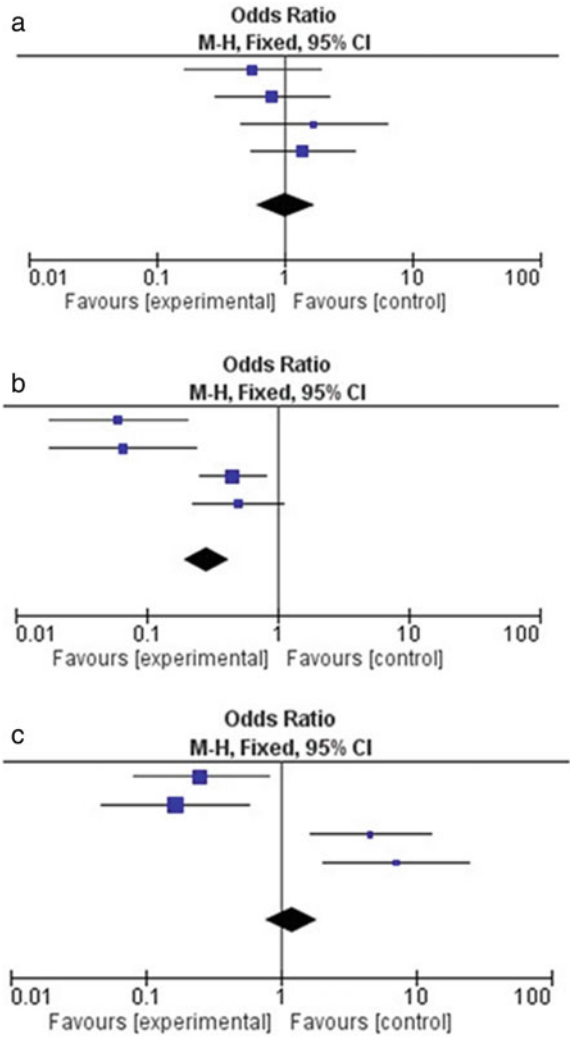

Fig. 3 Hypothetical forest plots—judgement regarding inconsistency

estimate remains intact; hence, we should not downgrade for inconsistency. In the forest plot B, all the four trials have the same direction of effect; however, the magnitudes of effect estimates vary, and there is little overlap between the confidence intervals between first and second two studies. Therefore, the forest plot shows inconsistency. However, the inconsistency probably does not decrease our confidence in the pooled estimate. Hence, we may not downgrade for inconsistency. In the forest plot C, the magnitude of difference in the effect estimates between the first two and second two studies is similar to that in the forest plot B, but the direction of effects are opposite. The first two studies favour intervention while the latter two studies favour control. Therefore, the

forest plot shows inconsistency. Does the inconsistency decrease the confidence in the pooled estimate? Probably yes, and hence, we should downgrade certainty of the evidence for inconsistency.

(2) *Statistical tests*. The two commonly used statistical tests for inconsistency (heterogeneity) are the Chi-squared test (test for heterogeneity) and the I^2 test. The Chi-squared test examines the null hypothesis that all studies evaluate the same effect. A p-value of < 0.05 for Chi-squared test indicates heterogeneity. I^2 test quantifies heterogeneity and can be used to compare heterogeneity across meta-analyses of different sizes, of different types of studies, and different types of outcome data (Higgins et al. 2003). I^2 value of < 40, 30–60, 50–90, 75–100% indicate low, moderate, substantial, and considerable heterogeneity respectively. The disadvantage of the I^2 test is that the cut-off values are not established, and judgement is required when the values fall in the overlapping zone. Chi-squared test and I^2 values are calculated in RevMan 5.4 software, and the values are displayed at the bottom of the forest plot (Review Manager. Version 5.4. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2020).

Suggestions for downgrading: The judgement regarding inconsistency requires careful evaluation of the forest plot and statistical tests. Downgrade for inconsistency only if it decreases our confidence in the pooled effect estimate.

Indirectness

Direct evidence comes from research that directly compares the interventions in which we are interested when applied to the populations in which we are interested, and measures outcomes important to patients (Guyatt et al. 2011). Indirectness refers to the extent to which the people, interventions, and outcome measures are different from those of interest. The fourth cause of indirectness results when there is no direct comparison between the two interventions of interest.

(1) ***Indirectness resulting from differences in the population of interest***: Systematic reviews will include only those studies which fulfil criteria with regards to population. However, indirectness can still result in some situations. For example, systematic review plans to investigate the effect of drug A in a patient population of individuals > 60 years. After performing a literature search, the reviewers notice that many studies examining drug A had 70 years or more eligibility criteria. In this case, the studies recruiting patients exclusively above 70 years of age still satisfy the inclusion criteria for the systematic review. Still, the age criteria of the included studies and the systematic review are not identical. Therefore, the effect of indirectness must be considered when concluding such situations. In this example, the reviewers may consider downgrading the level of evidence by one level if (a) there is a physiological

basis to assume that the effect of drug A in population >60 years is likely to be significantly different from the effect in a population exclusively >70 years of age, and (b) the studies with population exclusively >70 years' of age contribute a significant amount (weightage) of information to the pooled effect estimate.

(2) ***Indirectness resulting from differences in intervention***: Indirectness results when reviewers want to compare drug A to drug B; however, there is no direct comparison of drug A to drug B. Instead, the studies have compared drug A to drug C, and drug C to drug B. This type of indirectness is uncommon in systematic reviews. A more common reason for indirectness maybe when the studies have used only a part of rather than whole intervention. For example, a systematic review aims to investigate the effect of a group of interventions A-B-C-D for expediting post-operative recovery. The reviewers find that many studies have used only interventions A-C-D. The reviewers must consider the effect of indirectness if they include the studies with intervention A-C-D in the systematic review. The decision regarding downgrading certainty of evidence depends on whether the difference in the interventions (A-B-C-D versus A-C-D) is likely to have a significant effect on the outcome of interest (post-operative recovery) and amount of information (weightage) contributed by the studies with A-C-D intervention.

(3) ***Indirectness resulting from differences in the outcome***: This is a common reason for indirectness in systematic reviews. It may result for two reasons:

   (i) *Differences in the time frame*: For example, if the reviewers are interested in the intervention effect at 12 months but include studies that have reported effect only at six months. Suppose there is evidence that for other similar interventions, the effect decreases significantly from 6 months to 12 months, and a significant amount of information comes from the studies which have reported effect only until six months. In that case, the reviewers may decrease the level of certainty for indirectness.

   (ii) *Use of surrogate outcome*: Indirectness results when studies report only surrogate markers of the clinically meaningful outcomes; for example, HbA1c for symptoms of diabetes, C-reactive protein for sepsis. In such scenarios, reviewers should consider indirectness resulting from the difference in the outcome while grading the level of evidence

(4) ***Indirectness when there is no direct comparison between two interventions of interest***: This type of indirectness results when reviewers want to compare intervention A versus intervention B; however, the studies have compared intervention A versus intervention C and Intervention B versus intervention C. The indirect comparison requires assumption to be made that the population characteristics, co-interventions, outcome measurement, and the methodological qualities are not significantly different between the studies to result in different effects (Song et al. 2009). Because this assumption is always in some doubt, indirect comparisons always warrant rating down the quality of evidence by one level.

Suggestions for downgrading: The reviewers should consider rating down the certainty of evidence if indirectness is likely to influence the outcome of interest, and the significant amount of information comes from the studies with indirectness. Reviewers may rate down by one level when indirectness comes from a single factor, and by two if it comes from multiple factors. The decision requires judgement and consideration of the overall impact of the indirectness on the effect estimate.

Imprecision

Precision refers to the degree of agreement between repeated measurements. If repeated measures are close together, our confidence in the results increases as they are more likely to be close to the real population value. Thus precision is a surrogate marker of accuracy. The judgement regarding precision is based on 95% confidence intervals and sample size.

(1) *Confidence intervals:* Confidence intervals represent a range of values based on sample data, in which the population value is likely to lie. Confidence intervals are the measure of the precision of a mean. In general, for systematic reviews, precision is adequate if 95% confidence intervals exclude no effect.

In hypothetical forest plots (Fig. 4) of two systematic reviews A and B, the confidence interval does not cross the line of no effect in the forest plot A indicating "no imprecision". In contrast, it crosses the line of no effect in systematic review B, indicating "imprecision".

(2) *Optimal information size*: The results of a systematic review are reliable only when the confounding factors which influence the outcome are balanced between the intervention and control groups. The confounding factors to be balanced between the two groups require a minimal number of patients, often referred to as "Optimal information size", randomised to either intervention or control group. Optimal information size equals to the number of patients required to conduct an adequately powered RCT.

The importance of fulfilling criteria for optimal information size is evident from the following example: A systematic review and meta-analysis compared intravenous magnesium versus placebo in patients with suspected myocardial infarction for prevention of death (Fig. 5) (Teo et al. 1991; Guyatt et al. 2011). The meta-analysis showed a significant beneficial effect of the intervention with an odds ratio of 0.44 and confidence intervals 0.27 to 0.71. Even though the effect estimate's confidence interval did not cross the line of no effect, one may not be confident in the results because of the small sample size and fewer events. In such

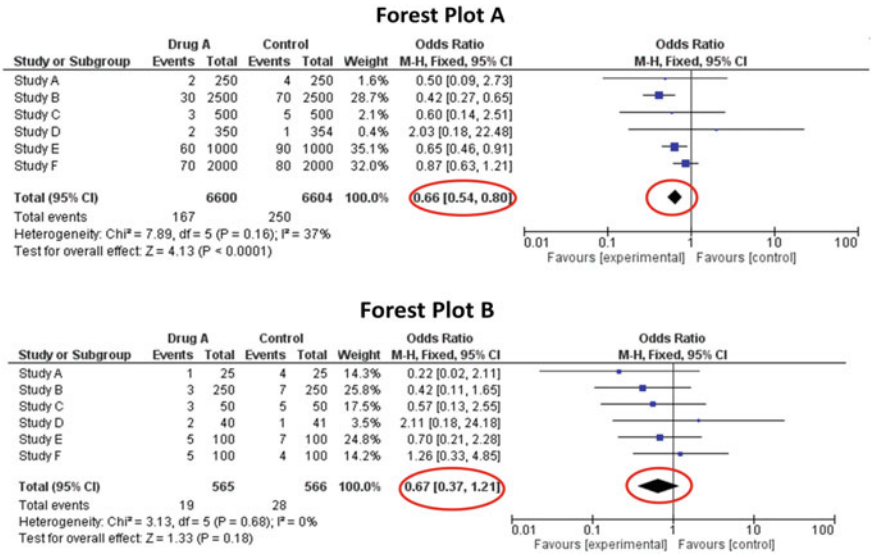

Fig. 4 Hypothetical forest plots—judgement regarding imprecision

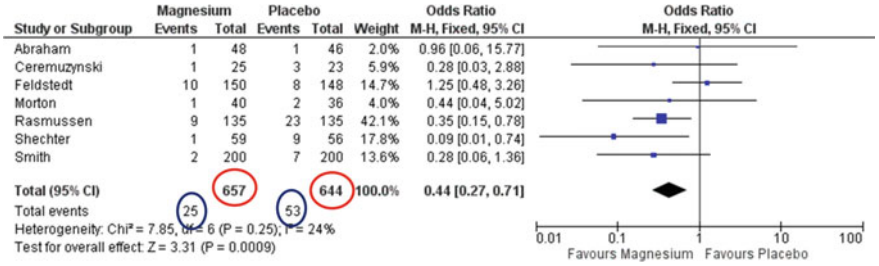

Fig. 5 Forrest plot comparing intravenous magnesium versus placebo in patients with suspected myocardial infarction for prevention of death (Teo 1991)

situations, it may be reasonable to downgrade the certainty of the evidence for imprecision because of small information size.

Suggestions for downgrading: Do not downgrade for imprecision if optimal information size criterion is met, and confidence interval excludes no effect (i.e., relative risk (RR) of 1.0). Downgrade by one level if optimal information size criterion is not met or if the confidence interval fails to exclude significant benefit or harm (e.g., overlaps RR of 1.0). Reviewers may consider rating down by two if both the criteria (Confidence interval and optimal information size) are not met or when the confidence interval is very wide (Guyatt et al. 2011).

Publication/Reporting Bias

Publication bias is a reporting bias that results from failure to identify all relevant trials. Publication bias occurs from the publication or non-publication of relevant trials, depending on the nature and direction of the results (Sedgwick 2015). Trials with positive findings are more likely to be published than trials with negative or null findings (RR 1.78, CI 1.58 to 1.95) (Hopewell et al. 2009). Therefore, a meta-analysis in the presence of publication bias is likely to over-estimate the treatment effect. If a systematic review contains studies predominantly with small sample sizes or studies sponsored by the pharmaceutical industry, it increases publication bias. The pharmaceutical sector discourages publication of trials they supported, which have negative findings (Egger and Smith 1998).

The other sources of reporting bias include time-lag bias (delay in the publication of trials with negative findings), language bias (not including studies published in languages other than English), and bias arising from publication of trial in "grey literature" (e.g., theses, conference abstracts, un-indexed journals). These sources of bias prevent an eligible study from being identified and included in the systematic review.

The presence of reporting bias in a systematic review is assessed by visual inspection of the funnel plot for symmetry and Egger's test. Funnel plots are scatter-plots of the studies in a meta-analysis, with the treatment effect on the horizontal axis and some measure of weight, such as the inverse variance, the standard error, or the sample size, on the vertical axis (Lau et al. 2006). Generally, effect estimates from large studies will be more precise and will be near the apex of an imaginary funnel. In contrast, results from smaller studies will be less precise and would lie towards the funnel base evenly distributed around the vertical axis. Asymmetric distribution of the studies around the vertical axis raises the possibility of publication bias. However, apart from publication bias, a skewed funnel plot may result from other causes: by chance, true heterogeneity in the intervention effect, and statistics used to measure effect size.

Suggestions for downgrading: Consider rating down the evidence if the evidence is based mainly on multiple small trials, especially when industry-sponsored or investigators have conflicts of interest. Consider rating down the evidence when publication bias is strongly suspected based on funnel plot asymmetry. As there is no full-proof method to prove or rule out publication bias or to determine a threshold for publication bias, GRADE suggests systematic reviewers to decide whether publication bias was "undetected" or "strongly suspected" in a systematic review. Because of the uncertainty in assessing the likelihood of publication bias, GRADE suggests rating down by a maximum of one level when publication bias is strongly suspected (Guyatt et al. 2011).

Factors that can improve the certainty of evidence in systematic reviews of observational studies: Generally, evidence generated from observational studies is considered as "Low" certainty. However, in the following rare circumstances, observational studies can produce moderate or high certainty evidence.

(1) When methodologically robust observational studies show large or very large and consistent treatment effect, the treatment and effect relationship is likely to be stronger. In these situations, the reviewers may consider upgrading the certainty of evidence by one level.
(2) When studies show a dose-response effect, the effect is more likely related to the intervention. Hence the reviewers may consider upgrading certainty of evidence by one level.
(3) When plausible biases or confounding factors are likely to decrease the effects of an intervention, reviewers may consider upgrading the certainty of evidence by one level.

Summary of Findings Table

A Summary of Findings (SoF) table summarises the critical results of a systematic review. It also informs the readers about the level of reviewer's confidence in the results based on the GRADE approach. The SoF table allows reviewers to make explicit judgements regarding the certainty of evidence and readers to understand the reasoning behind the judgements. The GRADEpro Guideline Development Tool (GRADEpro GDT) is online software (available at https://gradepro.org/) used to create a summary of findings table for systematic reviews.

Summary

GRADE offers a system for rating certainty of evidence in systematic reviews. In this chapter, we have discussed the critical aspects that systematic review authors need to consider while grading the certainty of evidence. The GRADE process requires judgement and is not objective, but it does provide a transparent and well-defined method for developing and presenting evidence summaries for systematic reviews.

References

Atkins D, Best D, Briss PA, Eccles M, Falck-Ytter Y, Flottorp S, et al. Grading quality of evidence and strength of recommendations. BMJ. 2004;328(7454):1490.
Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol. 2011;64(4):401–406.
Egger M, Smith GD. Bias in location and selection of studies. BMJ. 1998;316(7124):61–6.
Guyatt GH, Oxman AD, Kunz R, Brozek J, Alonso-Coello P, Rind D, et al. GRADE guidelines 6. Rating the quality of evidence–imprecision. J Clin Epidemiol. 2011;64(12):1283–1293.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence–inconsistency. J Clin Epidemiol. 2011;64(12):1294–1302.

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence–indirectness. J Clin Epidemiol. 2011;64(12):1303–1310.

Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, et al. GRADE guidelines: 5. Rating the quality of evidence–publication bias. J Clin Epidemiol. 2011;64(12):1277–1282.

Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, et al. GRADE guidelines: 4. Rating the quality of evidence–study limitations (risk of bias). J Clin Epidemiol. 2011;64 (4):407–415.

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650):924–6.

Higgins JPT SJ, Page MJ, Elbers RG, Sterne JAC. Chapter 8: Assessing risk of bias in a randomised trial. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook of systematic reviews of interventions, 2nd edn. Chichester (UK): John Wiley and Sons; 2019. p. 205–228.

Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327(7414):557–60.

Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009; (1): Mr000006.

Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. BMJ. 2006;333(7568):597–600.

Puhan MA, Schünemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ. 2014;349:g5630.

Schunemann HJ HJ, Vist GE, Glasziou P, Akl EA, Skoetz N, Guyatt GH. Chapter 14: completing 'summary of findings' tables and grading the certainty of the evidence. In: Higgins JPT TJ, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. Cochrane handbook for systematic reviews of interventions, 2nd edn. Chichester (UK): John Wiley and Sons; 2019. p. 375–402.

Sedgwick P. What is publication bias in a meta-analysis? BMJ. 2015;351:h4419.

Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ. 2009;338:b1147.

Teo KK, Yusuf S, Collins R, Held PH, Peto R. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. BMJ. 1991;303(6816):1499–503.