Rating the certainty in evidence in the absence of a single estimate of effect

M Hassan Murad,[1] Reem A Mustafa,[2,3] Holger J Schünemann,[3] Shahnaz Sultan,[4] Nancy Santesso[3]

10.1136/ebmed-2017-110668

[1]Evidence-based Practice Center, Mayo Clinic, Rochester, Minnesota, USA
[2]Department of Biomedical and Health Informatics, University of Missouri-Kansas City, Kansas City, Missouri, USA
[3]Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, Ontario, Canada
[4]Division of Gastroenterology, Department of Medicine, University of Minnesota, and Minneapolis Veterans Affairs Health Care System, Minneapolis, Minnesota, USA

Correspondence to
Dr M Hassan Murad,
Evidence-based Practice Center, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA;
murad.mohammad@mayo.edu

Abstract

When studies measure or report outcomes differently, it may not be feasible to pool data across studies to generate a single effect estimate (ie, perform meta-analysis). Instead, only a narrative summary of the effect across different studies might be available. Regardless of whether a single pooled effect estimate is generated or whether data are summarised narratively, decision makers need to know the certainty in the evidence in order to make informed decisions. In this guide, we illustrate how to apply the constructs of the GRADE (Grading of Recommendation, Assessment, Development and Evaluation) approach to assess the certainty in evidence when a meta-analysis has not been performed and data were summarised narratively.

Background

Practitioners of evidence-based medicine need to know the level of certainty in the evidence they are applying to patient care. Whether they are using recommendations from a clinical practice guideline based on a systematic review of the literature or using the results directly from a systematic review, they need to know how trustworthy the evidence is with regard to the benefits and harms of a treatment or a diagnostic test. This construct is called certainty or quality of evidence. The GRADE (Grading of Recommendation, Assessment, Development and Evaluation) approach is a modern framework for rating the certainty in evidence.[1] Using GRADE, randomised controlled trials and observational studies are considered to generate high and low certainty evidence, respectively. This initial grade that is based on study design is modified using several key domains such as the methodological limitations of the studies, indirectness of the evidence to the question at hand, imprecision of estimates, inconsistency of the evidence, and the likelihood of publication bias. A body of evidence about a specific outcome is downgraded or upgraded to a final rating of high, moderate, low or very low. High certainty in evidence means that the investigators are very confident that the effect they found across studies is close to the true effect, and very low means that they have very little confidence in the effect.[1]

Often, a single pooled effect estimate from a meta-analysis is available and is used for assessing the certainty in evidence. However, when studies measure outcomes differently or report outcomes in ways that cannot be standardised and meta-analysed, or in situations of urgency, only a narrative synthesis might be available. Consider a systematic review of self-management programmes in patients with chronic obstructive pulmonary disease.[2] There were five randomised trials that informed the effect of the intervention on respiratory symptoms. The individual studies presented their results using different tools and measures which precluded pooling. Two trials[3 4] used the Borg scale to assess respiratory symptoms. One of these trials[3] also presented results for respiratory symptom severity. The third trial[5] presented results as the proportion of days rated in patients' diaries as having mild, moderate or severe respiratory symptoms. The fourth trial[6] presented results about mean breathlessness and sputum production scores over 2-week periods and the fifth trial[7] presented results as breathlessness, sputum volume and sputum colour during exacerbations. These studies could not be pooled, but the evidence could be summarised narratively. There is some guidance about how to synthesise the effects of interventions narratively.[8] Guidance about how to grade certainty in this evidence is needed. A judgement on the certainty in evidence is still required because certainty is a key component of decision-making. Providing decision makers (patients, clinicians and policymakers) with evidence of unknown trustworthiness compromises their ability to transform evidence to action.[9] Decision makers would want to know how confident we are in the effect of these programmes to improve respiratory symptoms before offering such programmes to patients with chronic obstructive pulmonary disease.

The approach

We provide suggestions on the use of GRADE to rate the certainty of evidence when a meta-analysis has not been performed, and instead a narrative summary of the effect was provided. The approach leverages the meaning of the constructs that represent GRADE domains to produce judgements on how these constructs affect our certainty. In table 1, we explain how the GRADE domains (methodological limitations of the studies or risk of bias, indirectness, imprecision, inconsistency and the likelihood of publication bias) can be applied without a single pooled estimate. Note that this guidance does not address meta-narrative reviews[10–13] (which answer questions about conceptual underpinnings and understanding of a phenomenon) or qualitative systematic reviews[14] (which summarise themes from focus groups and interviews); rather, we address evidence synthesis of quantitative estimates of effect not amenable to meta-analysis (and thus summarised narratively).

Example

In table 2, we again refer to the systematic review of self-management programmes in patients with chronic obstructive pulmonary disease[2] and illustrate how we applied the GRADE approach. The outcome of interest in this table is respiratory symptoms which were not pooled in meta-analysis. Evidence derived from five randomised trials showed small to no reductions in respiratory symptoms and was judged to warrant low certainty (rated down for methodological limitations of the included studies and

Table 1 Applying the GRADE approach when evidence for an effect is summarised narratively (a meta-analysis is not available)

GRADE domain	How to apply the GRADE domain to evidence that has been summarised narratively
Methodological limitations of the studies	Make a judgement on the risk of bias across studies for an individual outcome. A sensitivity analysis is not possible to determine if the effect changes when studies at high risk of bias are excluded. It is possible to consider the size of a study, its risk of bias and the impact it would have on the summarised effect.
Indirectness	Make a global judgement on how dissimilar the research evidence is to the clinical question at hand (in terms of population, interventions and outcomes across studies).
Imprecision	Consider the optimal information size (or the total number of events for binary outcomes and the number of participants in continuous outcomes) across all studies. A threshold of 400 or less is concerning for imprecision.[15] Results may also be imprecise when the CIs of all the studies or of the largest studies include no effect and clinically meaningful benefits or harms.
Inconsistency	Judge inconsistency by evaluating the consistency of the direction and primarily the difference in the magnitude of effects across studies (since statistical measures of heterogeneity are not available). Widely differing estimates of the effects indicate inconsistency.
Likelihood of publication bias	Publication bias can be suspected when the body of evidence consists of only small positive studies or when studies are reported in trial registries but not published. Statistical evaluation of publication bias is not possible in this case. Publication bias is more likely if the search of the systematic review is not comprehensive.
Factors that can raise certainty in evidence: ▶ Large effect ▶ Dose–response gradient ▶ Plausible confounders or other biases increase the certainty in the effect	If one of the three domains that can increase certainty in a body of evidence (typically from non-randomised studies) is noted, consider rating up the grade of certainty, particularly if it is noted in the majority of studies.

Table 2 Illustrative example of rating the certainty in evidence in the absence of a single estimate of effect

GRADE domain	Judgement	Concerns about certainty domains
Methodological limitations of the studies	One out of five trials[7] had low risk of bias in the three items assessed (sequence generation, allocation concealment and blinding) but it was the smallest study (46 participants). Two other trials (56 and 129 participants)[3 5] did not report on any of the risk of bias items; making judgements not possible, which was concerning. The remaining two trials[4 6] (235 and 157 participants) explicitly reported lack of blinding, unclear sequence generation and allocation concealment. Therefore, we judged the trials to have serious methodological limitations.	Serious
Indirectness	The patients, intervention and comparators in the studies all provide direct evidence to the clinical question at hand. All interventions included an educational component (with some variation in the direct respiratory therapy component). The type and severity of the symptoms (outcome) was assessed using different scales in different trials. We judged the evidence to have no serious indirectness but noted some variability in the intervention and outcome measure.	Not serious
Imprecision	The total number of patients included in all the trials was ~600. Some trials reported small reductions, and other trials reported 'non-significant results' likely because of enrolling a small number of participants which resulted in wide CIs that included meaningful benefits and no effects. We judged the evidence to have borderline imprecision.	Not serious, borderline
Inconsistency	The direction and magnitude of effect varied across the different trials. Overall the results showed either small reduction in symptoms or no change. Two trials,[3 4] showed a small effect on dyspnoea at the 5% level using the Borg scale in favour of self-management education programme. In the third trial,[5] they found no significant between-group differences in the proportion of days rated as mild, moderate or severe in their respiratory status in symptom diaries. In the fourth trial,[6] no significant between-group differences were seen in mean breathlessness and sputum production scores over 2-week periods. However, small statistically significant differences in mean cough and sputum colour scores were seen in favour of the intervention group. In the fifth trial,[7] no significant differences were found between the scores of the intervention and control group during exacerbations (breathlessness, sputum volume and sputum colour). We judged the evidence to have serious inconsistency.	Serious
Publication bias	We did not strongly suspect publication bias because both negative and positive trials were published, and the search for studies was comprehensive.	Not suspected

The outcome of interest is respiratory symptoms. Data are derived from a systematic review of self-management programmes in patients with chronic obstructive pulmonary disease.

Table 3 Illustrative example of how the summary of findings can be presented to guideline developers

Outcome	Effect	Number of participants (studies)	Certainty in the evidence*
Respiratory symptoms Assessed using a variety of scales	Most studies showed small reductions in symptoms or no effect.	623 (5 randomised trials)	LOW†‡ ⊕⊕OO (due to serious risk of bias and imprecision)

The outcome of interest is respiratory symptoms (for which a single pooled effect estimate was not available and only a narrative synthesis of the evidence was provided).
*Commonly used symbols to describe certainty in evidence in evidence profiles: high certainty ⊕⊕⊕⊕, moderate certainty ⊕⊕⊕O, low certainty ⊕⊕OO and very low certainty ⊕OOO.
†Serious risk of bias across studies because of unclear or inadequate blinding, sequence generation and allocation concealment.
‡Serious imprecision and inconsistency were considered together as there were small effects, or 'no effects' reported in studies (likely due to wide CIs).

inconsistency). Based on this assessment, decision makers can conclude that self-management programmes may slightly reduce respiratory symptoms. This evidence could also be presented to decision makers in a summary of findings table (typically used in guideline development and generated using GRADEpro which allows narrative summaries of the evidence; https://gradepro.org). Table 3 shows one row of a summary of findings table with explanatory notes. The certainty of evidence in table 3 summarises the GRADE judgements about the different domains (all detailed in table 2) that collectively determined the certainty in evidence for one outcome (respiratory symptoms).

Discussion

Evidence-based practice is founded on making decisions using the best available evidence, whether it is based on a pooled single effect estimate, or on a narrative review of the individual studies informing each outcome. Stakeholders require that such evidence is appraised and the certainty in the effect is determined in order to inform decision-making. One of the greatest strengths of the GRADE approach is that it provides a systematic method to assess the certainty in evidence and a transparent documentation of the judgements used to assess the body of evidence. While typically it is thought to only apply to results that have been statistically aggregated, evaluating the certainty of evidence can also be performed when results have been narratively summarised.[16] In this setting, some certainty domains can be applied directly. For other domains, we have provided additional guidance in which the meaning and connotation of those domains can be used. Taken together, an overall assessment of the evidence can be determined. Stakeholders engaged in shared decision-making in a patient–physician dyad, in guidelines development, or in public health and policy, can then use the summarised effect and the certainty in the evidence to make informed decisions.

Competing interests None declared.

Provenance and peer review Not commissioned; internally peer reviewed.

References

1. **Balshem H**, Helfand M, Schunemann HJ, *et al*. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;**64**:401–6.
2. **Effing T**, Monninkhof EM, van der Valk PD, *et al*. Self-management education for patients with chronic obstructive pulmonary disease. *Cochrane Database Syst Rev* 2007;(4):CD002990.
3. **Gourley GA**, Portner TS, Gourley DR, *et al*. Humanistic outcomes in the hypertension and COPD arms of a multicenter outcomes study. *J Am Pharm Assoc (Wash)* 1998;**38**:586–97.
4. **Boxall AM**, Barclay L, Sayers A, *et al*. Managing chronic obstructive pulmonary disease in the community. A randomized controlled trial of home-based pulmonary rehabilitation for elderly housebound patients. *J Cardiopulm Rehabil* 2005;**25**:378–85.
5. **Watson PB**, Town GI, Holbrook N, *et al*. Evaluation of a self-management plan for chronic obstructive pulmonary disease. *Eur Respir J* 1997;**10**:1267–71.
6. **Monninkhof E**, van der Valk P, Schermer T, *et al*. Economic evaluation of a comprehensive self-management programme in patients with moderate to severe chronic obstructive pulmonary disease. *Chron Respir Dis* 2004;**1**:7–16.
7. **Bourbeau J**, Julien M, Maltais F, *et al*. Reduction of hospital utilization in patients with chronic obstructive pulmonary disease: a disease-specific self-management intervention. *Arch Intern Med* 2003;**163**:585–91.
8. **Popay J**, Roberts H, Sowden A, *et al*. Guidance on the conduct of narrative synthesis in systematic reviews: a product from the ESRC Methods Programme. 2006. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.178.3100&rep=rep1&type=pdf (accessed 11 Jan 2017).
9. **Alonso-Coello P**, Schunemann HJ, Moberg J, *et al*. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ* 2016;**353**:i2016.
10. **Abu Dabrh AM**, Firwana B, Cowl CT, *et al*. Health assessment of commercial drivers: a meta-narrative systematic review. *BMJ Open* 2014;**4**:e003434.
11. **Domecq JP**, Prutsky G, Elraiyah T, *et al*. Patient engagement in research: a systematic review. *BMC Health Serv Res* 2014;**14**:89.
12. **Mohammed K**, Nolan MB, Rajjo T, *et al*. Creating a patient-centered health care delivery system: a systematic review of health care quality from the patient perspective. *Am J Med Qual* 2016;**31**:12–21.
13. **Wong G**, Greenhalgh T, Westhorp G, *et al*. RAMESES publication standards: meta-narrative reviews. *BMC Med* 2013;**11**:20.
14. **Lewin S**, Glenton C, Munthe-Kaas H, *et al*. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS Med* 2015;**12**:e1001895.
15. **Guyatt GH**, Oxman AD, Kunz R, *et al*. GRADE guidelines 6. Rating the quality of evidence—imprecision. *J Clin Epidemiol* 2011;**64**:1283–93.
16. Thayer KA, Schünemann HJ. Using GRADE to respond to health threats with different levels of urgency. *Environment International* 2016;92-93:585–9.