Journal of Clinical Epidemiology 111 (2019) 83–93

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

GRADE guidelines: 20. Assessing the certainty of evidence in the importance of outcomes or values and preferences—inconsistency, imprecision, and other domains

Yuan Zhang[a], Pablo Alonso Coello[a,b,*], Gordon H. Guyatt[a], Juan Jose Yepes-Nuñez[a], Elie A. Akl[a,c], Glen Hazlewood[d], Hector Pardo-Hernandez[b], Itziar Etxeandia-Ikobaltzeta[a], Amir Qaseem[e], John W. Williams Jr.[f], Peter Tugwell[g], Signe Flottorp[h,i], Yaping Chang[a], Yuqing Zhang[a], Reem A. Mustafa[a,j], María Ximena Rojas[k], Feng Xie[a,l], Holger J. Schünemann[a,m,*]

[a]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Canada*
[b]*Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain*
[c]*Department of Internal Medicine, Faculty of Medicine, American University of Beirut, Beirut, Lebanon*
[d]*Department of Medicine and Department of Community Health Sciences, University of Calgary, Calgary, Canada*
[e]*American College of Physicians, Philadelphia, PA, USA*
[f]*Center of Innovation for Health Services Research in Primary Care at the Durham Veterans Affairs Medical Center and Duke University, Durham, NC 27701, USA*
[g]*Department of Medicine, University of Ottawa, Ottawa, Canada*
[h]*Norwegian Institute of Public Health, Oslo, Norway*
[i]*Institute of Health and Society, University of Oslo, Oslo, Norway*
[j]*Division of Nephrology and Hypertension, Department of Internal Medicine, University of Kansas Medical Center, Kansas City, KS, USA*
[k]*Department of Clinical Epidemiology and Biostatistics, Pontificia Universidad Javeriana, Bogotá, Colombia*
[l]*Program for Health Economics and Outcome Measures (PHENOM), Hamilton, Canada*
[m]*Department of Medicine, McMaster University, Hamilton, Canada*

Accepted 3 May 2018; Published online 22 May 2018

Abstract

Objective: To provide Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) guidance for assessing inconsistency, imprecision, and other domains for the certainty of evidence about the relative importance of outcomes.

Study Design and Setting: We applied the GRADE domains to rate the certainty of evidence in the importance of outcomes to several systematic reviews, iteratively reviewed draft guidance, and consulted GRADE members and other stakeholders for feedback.

Results: We describe the rationale for considering the remaining GRADE domains when rating the certainty in a body of evidence for the relative importance of outcomes. As meta-analyses are not common in this context, inconsistency and imprecision assessments are challenging. Furthermore, confusion exists about inconsistency, imprecision, and true variability in the relative importance of outcomes. To clarify this issue, we suggest that the true variability is neither equivalent to inconsistency nor imprecision. Specifically, inconsistency arises from population, intervention, comparison and outcome and methodological elements that should be explored and, if possible, explained. The width of the confidence interval and sample size inform judgments about imprecision. We also provide suggestions on how to detect publication bias and discuss the domains to rate up the certainty.

Ethics approval and consent to participate: Not required. This study does not involve de novo patient data collection. No patient informed consent and institutional review board approval have been sought.

Consent for publication: Not applicable.

Availability of data and materials: The data sets supporting the conclusions of this article are included within the article and its additional file.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. It was funded through internal research funds at McMaster University available to HJS. GH is supported by a CIHR New Investigator Salary Award and a The Arthritis Society Young Investigator Salary Award, neither of which is directly related to this research project.

Conflict of interest: All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/conflicts-of-interest/); no financial relationships with any organizations that might have an interest in the submitted work in the previous 3 years, no other relationships or activities that could appear to have influenced the submitted work. Authors are members of the GRADE Working Group.

* Corresponding authors. Department of Health Research Methods, Evidence, and Impact, McMaster University Health Sciences Centre, Room 2C16, 1280 Main Street West, Hamilton, ON L8N 4K1, Canada. Tel.: +1 905 525 9140 × 24931; fax: +1-905-522-9507.

E-mail address: schuneh@mcmaster.ca (H.J. Schünemann).

https://doi.org/10.1016/j.jclinepi.2018.05.011
0895-4356/© 2018 Published by Elsevier Inc.

Conclusion: We provide guidance and examples for rating inconsistency, imprecision, and other domains for a body of evidence describing the relative importance of outcomes. © 2018 Published by Elsevier Inc.

Keywords: GRADE; Quality of evidence; Importance of outcomes; Value and preference; Inconsistency; Imprecision; Publication bias

1. Introduction

The Grading of Recommendations, Assessment, Development and Evaluation (GRADE) working group developed a widely accepted approach to rate the certainty of a body of evidence in the contexts of systematic reviews, health technology assessment, health-care recommendations, and decision support [1−4]. This is the 20th in the ongoing series of articles describing the GRADE approach in the Journal of Clinical Epidemiology and complements articles in this and other journals. We previously described the reasons for decreasing and increasing the certainty of a body of evidence; how an overall rating of the evidence is performed [5]; how to create GRADE evidence profiles and summary of findings tables [6−9]; how evidence is used to move to recommendations and decisions [10−15]; how evidence is dealt with in particular circumstances of diagnostic, prognostic, equity-related, multiple treatment comparison, environmental and public health questions [16−21]; how GRADE applies to rapid advice [22]; how GRADE deals with new risk of bias tools [23]; and when there is missing outcome data [24].

Decisions in health care require an assessment of the certainty of how much people value the importance of the outcomes that researchers and practitioners intend to affect [10−15]. Indeed, a variety of instruments are in use to elicit the relative importance of outcomes, including health state value or utility, or willingness to pay instruments. Our previous article in this series introduced the topic and the typical instruments in some detail [26]. For example, patients with severe or very severe chronic obstructive pulmonary disease (COPD) are willing to pay $13.46 to avoid mild side effects and $67.51 to gain symptom relief [25]. This suggests patients place more value on avoiding no symptom relief than avoiding mild side effects. We also described the terminology regarding the relative importance of outcomes and how it relates to the concept of values and preferences. The term outcome includes ‘‘health state’’ and non-health states that are related to the interventions under consideration, a broad set of the outcomes directly and indirectly related to health or a disease or non−health consequences. We then introduced the GRADE approach to rate the certainty of research evidence that focuses on this gap area, the rating of the certainty of a body of evidence about the relative importance of outcomes with a focus on risk of bias and indirectness [26]. We also introduced our rationale and explanations about the terminology of outcome importance [26]. In this article, we will provide guidance on rating of a body of evidence about the relative importance of outcomes dealing with the GRADE domains' inconsistency and imprecision and describe other concepts related to publication bias, rating up the certainty of evidence and variability in estimates. This guidance informs the certainty in the evidence about the relative importance of the outcome (values) and, thus, the evidence to decision framework. However, it will also affect the assessment of the balance between desirable and undesirable health outcomes and eventually affect the balance of the overall desirable and undesirable consequences (including the consequences on other criteria in the evidence to decision framework, for example, ethical consequences or consequences on health equity). In addition, our approach will be of use for those using utilities and values and preferences apart from GRADE evidence to decision frameworks; in particular, those conducting health technology assessments and decision modeling.

2. Methodology

We described the detailed methods for this work in the previous article [26]. Briefly, we used an iterative multipronged approach to develop guidance for assessing the certainty of a body of evidence addressing the relative importance of outcomes. We applied the same GRADE domains (risk of bias, inconsistency, indirectness, imprecision, publication bias, and domains to rate up the evidence) to the evidence describing the relative importance of outcome ratings systematic reviews and developed guidance based on these examples [5]. We refined this guidance after group discussions of GRADE project group meetings and consultations with stakeholders for feedback.

3. Inconsistency

According to the GRADE approach, raters can lower the certainty of the evidence if there is unexplained inconsistency or heterogeneity. However, assessment of inconsistency of evidence about the relative importance of outcomes is challenging for several reasons. First, the existing systematic reviews or health technology assessments often lack a clear definition of the relative importance of outcomes or values and preferences and include a diverse set of methods and instruments to assess them [27−30]. Thus, it is often challenging to determine if observed differences in the relative importance of outcomes is due to instruments or other potential underlying factors. Second,

What is new?

Key findings

- We provide novel guidance for rating inconsistency, imprecision, and other domains for a body of evidence describing values and preferences, the relative importance of outcomes or utilities, the term often used in the context of health economic evaluations.

What this adds to what was known?

- Meta-analyses are uncommon for evidence about the relative importance of outcomes, which makes inconsistency and imprecision assessments challenging. This is because of the lack of statistical approaches to assess heterogeneity and challenges in estimating pooled confidence intervals (CIs). Inconsistency in this type of evidence arises from population, intervention, comparison, and outcome and methodological elements. Unexplained inconsistency leads to downgrading the certainty of evidence. Similar to the conventional Grading of Recommendations, Assessment, Development and Evaluation (GRADE) assessment, the width of the CI and the sample size inform judgments about imprecision.
- Variability in the estimates of relative outcome importance is a separate issue in relation to inconsistency and imprecision. Variability in the relative importance of outcome often will demand weak recommendations in guideline development.

What is the implication and what should change now?

- A body of evidence addressing the importance of outcomes starts at "high certainty"; risk of bias, indirectness, inconsistency, imprecision, and publication bias can lead to rating down this evidence. Users of evidence about the relative importance of outcomes are encouraged to assess the certainty of evidence and provide feedback to the GRADE working group.

quantitative synthesis of relative importance of outcomes is uncommon because systematic review authors are hesitant to pool estimates obtained with different instruments such as the standard gamble, time trade-off or rating scales. This creates a dilemma for interpretation of systematic reviews as qualitative rather than quantitative syntheses and thus creates a challenge for assessing the inconsistency domain. In other situations where methods such as discrete choice, willingness to pay, rankings, or other scales are used, there is often only one single study available. The judgment about inconsistency is straightforward in the latter case because inconsistency does not exist in the context of single study evidence (a body of evidence based on one study will likely be rated down for one or more of the other GRADE domains). Although we suggest that raters attempt to statistically pool the relative importance of outcomes if appropriate, if no pooled estimates are available, the assessment of inconsistency still follows the same principles that we suggest below.

We propose raters examine inconsistency for evidence about the relative importance of outcomes in the following steps: (1) assess inconsistency; (2) explore reasons for inconsistency if the results across studies are inconsistent and not rating down if inconsistency can be explained (Fig. 1); and (3) discuss the credibility of subgroup differences if they are detected. We will begin by describing the signaling questions for rating inconsistency (see Fig. 1 and Appendix 1).

3.1. Signaling question: are the results across the included studies consistent?

The four items for assessing inconsistency in the results are similarity in point estimates, overlap in confidence intervals (CIs), statistical test for heterogeneity, and I^2 values. We suggest the evaluation of point estimates and CIs by visual inspection. If meta-analyses are available, the statistical test for heterogeneity and I^2 allow for quantitative estimates of heterogeneity [31,32]. An alternative for I^2, which depends on the size of the studies included, is s^2 [32]. If examination of the items aforementioned suggests no important inconsistency, raters label the corresponding domain as "not serious," and affirm the signaling question. Otherwise, raters should consider exploring the source.

3.2. Detailed exploration of inconsistency

When results are not consistent, raters explore inconsistency in the following ways (see Appendix 1). Raters should evaluate differences in the population, e.g., demographic characteristics such as gender, age, or cultural background, the options, e.g., the dose, duration, or administration of medication in direct choice experiments, and outcomes assessed, e.g., same outcome but different severities. Raters should be aware that these elements are not completely independent, e.g., when patients assess the relative importance of outcomes of their own health outcome, disease severity is an element both related to population and the outcome. As discussed in the indirectness section, for the relative importance of outcomes, the difference in options would be a signal to suggest potential difference in outcomes. The consideration of treatment options is more relevant when a study evaluates direct choice and infers the outcome importance accordingly.

Methodological inconsistency may arise from the risk of bias inherent in the study design, measurement

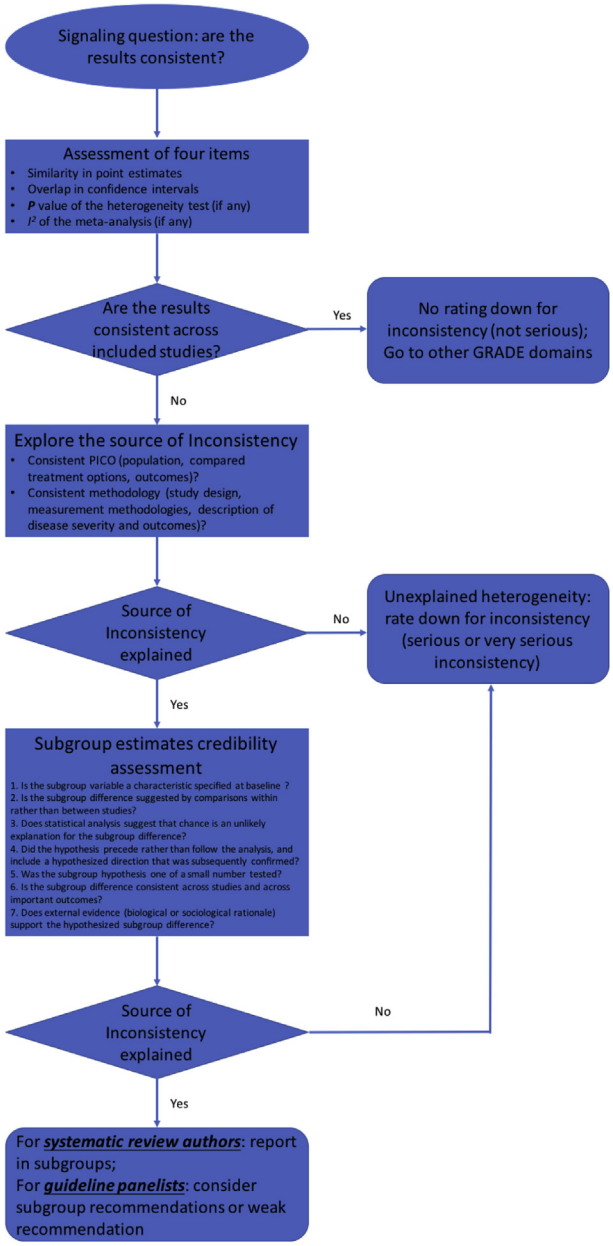

Fig. 1. Flow chart for assessment of inconsistency.

methodology (e.g., standard gamble or time trade-off, and visual analogue scale for utility), or description of outcomes or health states, such as narrative versus point by point format of health states, detailed versus less detailed descriptions. For example, inconsistency may be explained if different specific outcomes are measured such as those based on widely differing descriptions for severe stroke across included studies [33,34].

3.3. Credibility of subgroup estimates

Raters should formulate a priori hypotheses to explore inconsistency due to potential subgroup effects. If subgroup analyses show differences, raters should judge the credibility of the prespecified subgroup effects. While frameworks for evaluating subgroup effects of treatments exist [35], they do not exist for assessing the relative importance of outcomes. Until further guidance is available, we suggest raters make judgments on whether subgroup estimates are credible with the criteria proposed for subgroup effects of treatment [35].

3.4. Different strategies for systematic review authors and guideline panelists

Systematic review authors should only combine results from included studies if the results are similar enough. This can begin with pooling across studies and then test the assumption of similarity across studies. If there is substantial heterogeneity and systematic review authors discover that population, intervention, comparison and outcome (PICO) or methodological elements are a source of heterogeneity, they should narratively summarize and present the results for these groups of patients or people, compared alternatives, or outcomes. Guideline developers can then formulate recommendations separately for subgroups with different values or they can formulate a conditional or weak recommendation across populations indicating that how differences in values would affect implementation of the recommendation.

3.5. Example of inconsistency

A systematic review summarized the relative importance of psoriasis-related outcomes using the willingness to pay and utility approaches. Two included studies elicited willingness to pay for health states using the same instrument. However, important differences existed across these two studies in terms of willingness to pay for physical comfort ($2,000 vs. $10,000), social comfort ($1,000 vs. $2,000), emotional health ($2,000 vs. $5,000), self-care ($1,500 vs. $9,500), intimacy ($1,000 vs. $5,000), ability to sleep ($625 vs. $10,000), ability to work/volunteer ($1,600 vs. $10,000), and ability to concentrate ($875 vs. $7,500), respectively. The importance of the eight outcome measures differed across the two studies. For example, emotional health was one of the most important outcomes in one study ($2,000 and ranked top) but not in the other ($5,000 but ranked 6th) [36]. There were also no PICO elements that explained these differences, and one could justify rating down the certainty of evidence for serious inconsistency.

3.6. Variability versus inconsistency

When referring to inconsistency or heterogeneity across studies for the relative importance of outcomes, we suggest avoiding the term variability. True variability of the relative importance of outcomes within studies requires a separate assessment. We will discuss this issue further in the following.

4. Imprecision

Rating imprecision for the relative importance of outcomes includes an assessment of both the CI and sample size for the body of evidence. This assessment is often challenging because there are meta-analyses rarely and, thus, no calculated CIs. For the same reason, there is no simple way to calculate the minimum sample size to produce a sufficiently narrow estimate with sufficient power for the relative importance of the outcomes [37]. However, we suggest raters take the following approaches (see Fig. 2).

4.1. Confidence interval of the relative importance of outcomes

We recommend systematic review authors make their rationale for judgments explicit, such as accepting a certain range or assuming that decisions would be influenced or not influenced given the width of the CI. However, systematic review authors are often not in the best position to judge whether the CI around the estimate is sufficiently narrow for a specific decision (see Appendix 2). This is because this rating is usually dependent on the context including the type of interventions considered and resource expenditure. Furthermore, because of diversity in study designs, instruments, and presentation of results, CIs based on systematic reviews and meta-analyses may not be available. Under those circumstances, rating imprecision may be based on the number of studied people (sample size) alone.

For guideline panelists, we suggest rating imprecision based on whether the CIs of the relative importance of outcomes evidence cross a decision threshold (see Appendix 2). This requires taking particular absolute effect estimates of interventions on the outcomes into account for which the relative importance of outcomes is obtained. Imprecision is not present if the benefits clearly outweigh harms after combining the relative importance of outcomes and the absolute effect estimates, or vice versa, regardless of whether the upper or lower limit of the CI of the relative importance of outcomes estimate is assumed to be true in this calculation. If the decision would be overturned by assuming alternative estimates for the relative importance of outcomes stemming from the CI around them, raters should judge the evidence as seriously or very seriously imprecise.

4.2. Sample size

For both systematic review and guideline development, raters should also consider the sample size across studies when assessing imprecision. To assess imprecision, the review information size could be used as a threshold, and its calculation is likely to be different depending on the estimates used to indicate the relative importance of outcomes (e.g., utility, rank, or willingness to pay for an attribute) [38].

For guideline development with studies on direct choice and closely balanced benefits and harms that require a judgment about the direction of the recommendation (for or against an option), a potential approach is to use a threshold of 55% with a CI that exceeds the simple majority of 50% (i.e., that more than 55% of patients would make the same choice with an error margin just below 5%) [39,40]. For a single group, the sample size required to estimate this proportion (55%) with a confidence level of 95% and a desired precision of 5% (95% CI: 50–60%) would be 380 people [39]. Judgments about the GRADE imprecision domain for a body of evidence about the relative importance of outcomes may be informed by the rule of thumb if the body of evidence is based on at least 380 study participants. In situations when there is a potential large net benefit and one needs to decide about the strength of the recommendation (strong versus conditional) the GRADE Working Group suggests that 80% or 90% of people would make the same choice [39,40]. For evidence from direct choice studies, the sample size required to estimate these proportions (80% or 90%) with a confidence level of 95% and a desired precision of 5% would be 246 and 139 participants, respectively. Thus, using a "rule of thumb" approach in this context, raters may assess a body of evidence based on at least 250 patients or 140 patients, respectively, as precise.

In most situations, direct choice studies are not available, and for systematic reviews, judgments are made on a per outcome level. Under those circumstances, we suggest making a priori assessments of acceptable width of the CI for decision-making or using, again, a rule of thumb for sufficient precision of the relative importance of outcomes estimate. For example, a width of the CI of 0.1 on a utility scale, or margin of error, could be utilized to calculate the required review information size based on a confidence level (α), and population variance for the relative importance of outcomes.

4.3. Example of imprecision in the relative importance of outcomes

In a systematic review summarizing the utilities of COPD-related outcomes, one study measured the utilities of breathlessness using a visual analogue scale. The values across disease severities were 65.6 (standard error of the mean [SEM]: 3.4) for level 2 (stopping to catch breath after a few minutes walking) and 52.6 (SEM: 7.11) for level 3 breathlessness (breathless when dressing or washing) on a 0 to 100 scale, respectively. However, in this study, there were only 45 participants experiencing level 2 breathlessness and 7 participants level 3 breathlessness. Assuming the population standard deviation is 25, at the 95% confidence level, if the margin of error is 5 (that is to say, the width of CI is 10), we need at least 100 participants to estimate the population mean with sufficient precision. Thus, the small sample size did not meet the threshold for review information size. We downgraded one level for level 2 breathlessness and two levels for level 3 breathlessness for this concern about imprecision [41].

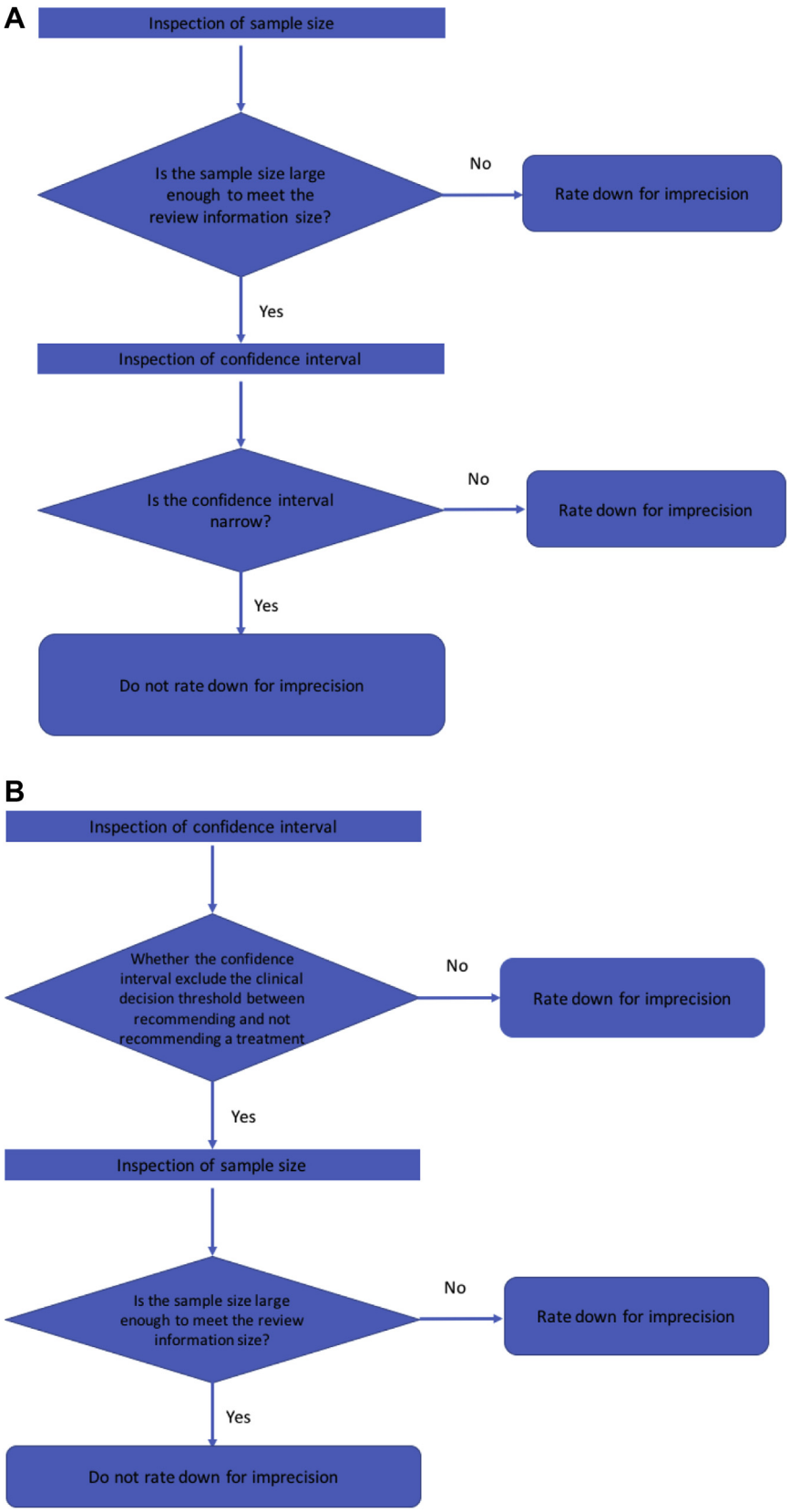

Fig. 2. Flow chart for assessment of imprecision. (A) Imprecision for systematic reviews. (B) Imprecision for guideline development.

5. Publication bias

Publication bias may be important for evidence addressing the relative importance of outcomes. Although the reasons for publication bias for this type of evidence may differ importantly from those of intervention studies where for-profit interest often play a role, other reasons for failure to publish (in a timely manner) may be similar. Conceivable reasons for delayed or unsuccessful publication include the results are not consistent with previous research results, results are redundant, or language or cultural circumstances lead to delays or failure to publish. Unfortunately, we are not aware of empirical evidence about the extent or how to properly assess publication bias in this field. Only in the situation that users have proof (through knowledge of conducted but unpublished studies) or strong suspicion of publication bias based on empirical knowledge, they should rate down the certainty of evidence for publication bias [42].

6. Rating up

The theoretical basis and empirical examples for using existing domains for rating up the certainty of evidence (a large effects or associations, dose-response gradient and direction of plausible residual confounding) of the relative importance of outcomes is limited. Thus far, we do not have clear guidance for when the evidence of the relative importance of outcomes should be rated up but we will describe some plausible scenarios here.

The certainty of a body of evidence summarizing the relative importance of outcomes starts as high certainty of the evidence. In other GRADE guidance for increasing the certainty of evidence, we suggest that raters only upgrade studies that are unlikely to be prone to bias or are imprecise [43]. The same considerations apply here. First, we suggest considering if any serious concerns about risk of bias or imprecision exist that would be so severe that they prohibit upgrading the certainty of evidence. If there are no such concerns or the concerns are limited on these or the other domains, the GRADE domains for increasing the certainty of evidence may apply. For instance, conceivable situations for what would be analogous to dose–effect relations are clear gradients in the relative importance of outcomes across different severity levels of marker states. Indeed, in a systematic review on how COPD patients value their outcomes, the pooled estimates for EQ-5D measurements of mild, moderate, severe, and very severe COPD are 0.85 (95% CI: 0.84–0.86), 0.80 (95% CI: 0.79–0.80), 0.72 (95% CI: 0.72–0.73), and 0.68 (95% CI: 0.67–0.69), respectively. Although we observed inconsistency for the utility values of COPD states across studies, we also identified a clear gradient of disutility as the disease progresses. This observation would increase our overall confidence that there is a gradient in utilities of the health states and mitigate concerns about the certainty of evidence for inconsistency. This is not suggesting the "gradient" as a consideration for inconsistency, but suggesting that in the context of other limitations, there may not be enough reason to rate down by one entire level [5,43]. Other plausible situations for rating up may arise when two health states differ importantly in their relative importance of outcomes, and, even if it were smaller than observed, the difference is sufficient to inform decision-making. That is, if the difference between outcomes is precise, and the studies not importantly biased, overall certainty in the difference may be solidified by large observed differences despite possible concerns about other domains such as indirectness. Raters could use the minimal important difference of the relative importance of outcomes such as 0.05 to 0.07 on a 0 to 1 visual analogue scale for making such judgments [44]. However, we need to further explore the application of the domains that increase certainty or mitigate concerns about certainty on existing domains. As we continue to develop the GRADE approach for evidence addressing the relative importance of outcomes, the current GRADE domains for rating up certainty of evidence may evolve or other reasons to rate up may arise.

7. Distribution (variability) of the relative importance of outcomes

7.1. Distribution across individuals and decision-making scenarios

We developed GRADE guidance for rating the certainty (or quality) of evidence. In this section, we will describe how it relates to decision-making and how it informs the GRADE EtD criterion "how much people value the main outcomes" [10]. Until we will have developed further guidance, we suggest to not rate the certainty in the variability of the relative importance of outcomes but making the potential for underlying variability transparent. The term "variability" of values is used ambiguously. It has been indistinctly used to refer to the inconsistency of results across studies (inconsistency), the width of the CIs (imprecision), or the distribution within a population. We suggest that these concepts be kept apart. We described our approach for addressing inconsistency and imprecision and the reasons for rating down for these domains mentioned previously. When we refer to variability, we mean biological variability (including psychological or cognitive) in the relative importance of outcomes for which there is no (current) explanation. For example, patients show a large degree of variability with regard to how they value the relative importance of gastrointestinal bleeds in the context of stroke prevention (Fig. 3). The reasons for why some patients are very averse to bleeding events and others only somewhat averse are not well understood but they eventually would have a biological, perhaps biopsychological, explanation. This variability in importance attached to specific outcomes can lead to both inconsistency in the results across studies (if patients with different although unknown predictors for how they rate the relative importance of outcomes are included) or

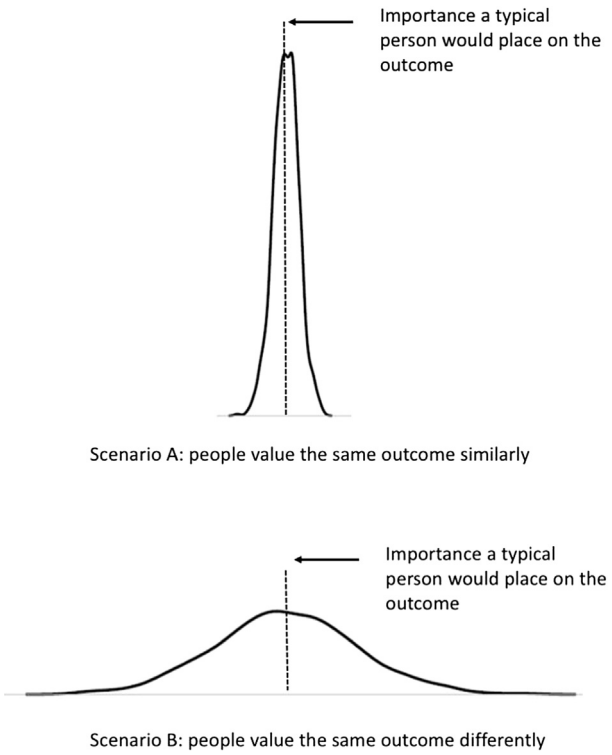

Fig. 3. Variability: the wide distribution of relative importance of outcome across individuals and/or decision-making scenarios.

imprecision (if there is a wide distribution of values for the relative importance of bleeding within studies). However, in the context of this variability, inconsistency may not apply when all studies include a large number of patients with different values for the relative importance, leading to wide but overlapping CIs. Imprecision may be the explanation if the sample size is small. However, imprecision may not apply if there are a large number of patients with different values for the relative importance of bleeding within studies. If the explanations of inconsistency or imprecision are not present, then true variability may exist.

7.2. Deciding about variability in the relative importance of outcomes

Judging whether the variability of the relative importance of outcomes is important or not requires balancing all outcomes and other GRADE EtD criteria. Indeed, variability in how patients value the main outcomes will influence the strength of a recommendation. Guideline panelists should consider whether the potential variability is important enough for them to make different recommendations across the patient subpopulations that value the outcomes differently or offer a conditional, value sensitive recommendation, across subpopulation emphasizing the need to elicit the relative importance of the outcome in the decision-making context carefully.

8. A practical example of assessing the overall certainty in the relative importance of an outcome across a body of evidence

We describe how the GRADE approach can be applied to assess the certainty of evidence for the relative importance of outcomes. The ratings start as "high" for all outcome assessments. Raters lower the certainty to moderate, low, or very low if one or more of the risk of bias, inconsistency, indirectness, imprecision or publication bias are judged as serious or very serious for the body of evidence, and consider the upgrading domains, to determine the final assessment of certainty of evidence (Fig. 4).

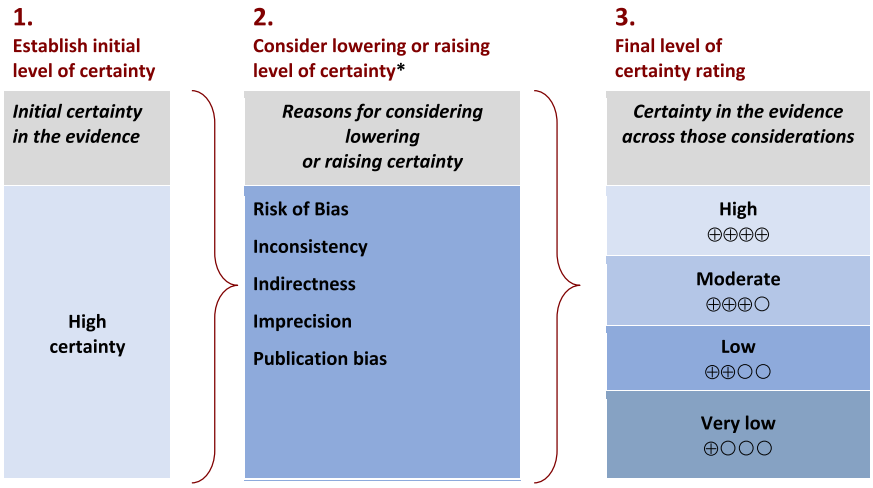

Fig. 4. Obtaining a final rating for the certainty of the evidence. * The presence of characteristics that increase confidence (gradient in disutility or utility as the outcome become severer, or large difference between outcomes weighed) may mitigate rating down for another domain. However, addressing domains for rating down quality of evidence (risk of bias, imprecision, inconsistency, indirectness, and publication bias) should take place in before considering reasons for rating up quality.

In the example shown in Table 1, we assessed the certainty of evidence for utility of nonfatal severe stroke. Seven studies on 580 participants were included [45–51], the pooled estimate was 0.15 (95% CI: 0.14–0.16) on a scale from 0 to 1. In some of the studies included, we had concerns about risk of bias due to either low response rate or failure to understand the utility instrument [46,49,51]. However, this only impacted a small proportion of the included study population, and the estimates based on low risk of bias studies were similar to those from studies subject to risk of bias. Therefore, we did not rate down the certainty of the body of evidence for risk of bias. The estimates in the included studies were similar. The pooled estimate was based on a large sample size, and the CI was narrow. Thus, we did not rate down for inconsistency or imprecision either. The included participants were taking antithrombotic treatment and at the risk of developing stroke or recurrent stroke, and they are the population of interest for answering the research question of "the relative importance of outcomes of interest in decision-making for patients with antithrombotic treatment." We were unable to identify anecdotal evidence suggesting publication bias. In summary, the certainty of evidence for this assessment is high, and we are very confident that severe stroke with its low utility has a large impact on lives and that severe stroke is a critical outcome to consider in decision-making.

9. Summary

This and the prior article in this series describe the GRADE approach for rating the certainty of evidence in the relative importance of outcomes or values and preferences [26]. Both the expansion of GRADE to this field of evidence and the assessment of a body of evidence in this area, in general, are innovative. The approach should be useful for systematic reviews, health technology assessment, decision modeling, and guidelines. The major challenge of rating the evidence about relative importance of outcomes is the widely differing research that exists in this field. Owing to this heterogeneity, evidence synthesis and, particularly, meta-analyses are uncommon. While these factors make the rating of the evidence challenging, the lack of pooled estimates is particularly problematic for rating inconsistency and imprecision. Better standardization of conduct and reporting of studies in this field should alleviate these challenges over time. This will have to be accompanied by further development of systematic review methodology for the relative importance of outcomes. Another challenge is that those summarizing and presenting evidence may not clearly separate the variability in how patients value main outcomes and mix them with the assessment of inconsistency and imprecision.

Despite all the challenges, we provide an explicit, structured, and transparent approach to assess the certainty of a body of evidence for the relative importance of outcomes. Health researchers, including systematic review authors, assessors of health technologies, and guideline developers, will now be in a position to assess the certainty of evidence for both the effects of interventions on outcomes and how important these outcomes are for the target populations (see Appendix 3 for more examples). This approach critically informs decision processes and provides key information for the "value" criterion in the GRADE EtD frameworks [10–15]. The approach will also allow

Table 1. Example of GRADE assessment for the certainty of evidence

Evidence profile								
Author(s): Yuan Zhang, Pablo Alonso Coello, Holger Schünemann Date: 2017-05-01								
Question: What are the views about the relative value/importance of outcomes of interest in decision-making for patients with antithrombotic treatment?								
Setting: not specified Bibliography: MacLean S. Chest 2012; 141:e1S-e23S.								
Quality assessment								
Outcome	**Study design/ measurement instrument**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other**	**Estimate of outcome importance (95% CI or other measure of variability)**	**Certainty**
Stroke								
Nonfatal severe stroke	7 studies, 580 participants VAS, SG, TTO	Not serious[a,b,c,d]	No serious inconsistency	No serious indirectness	No serious imprecision	None	0.10–0.39 (range of the point estimates) Pooled mean 0.149, 95% CI: 0.135–0.163	⊕⊕⊕⊕ High

CI, confidence interval; SG, Standard Gamble; TTO, time trade-off; VAS, visual analogue scale.

[a] The representativeness of the studies was impacted by a low response. However, this only impacted a small proportion of the included study population, and thus, we did not rate down for risk of bias.

[b] In Protheroe 2000, 97 of 260 invited patients responded.

[c] In Thomson 2000, 57 of the 180 invited patients completed the interview.

[d] 17.4% of participants in Gage 1995 did not understand the time trade-off technique.

assessing the certainty in a body of evidence informing (decision-analytical) models that rely on utilities which often are critical input parameters that may be utilized without scrutiny or based on single studies.

When balancing the desirable and undesirable effects of alternative options, we rate the overall certainty of the evidence, which requires combining the ratings from intervention effects and the relative importance of outcomes. While GRADE will have to elucidate how the overall certainty of evidence is expressed when ratings from intervention effects and the relative importance of outcomes are combined, we suggest that decision-makers consider both when balancing the desirable and undesirable consequences of alternative options [10−15].

Acknowledgments

The authors are grateful to Dr. Amiram Gafni from McMaster University for the comments on the manuscripts.

Authors' contributions: H.J.S. conceived of the project and approach; Y.Z., P.A., G.G., and H.J.S. designed the methodology for this project. Y.Z., J.J.Y.N., and Y.C. summarized the certainty assessment of relevant items in systematic reviews; Y.Z., P.A., G.G., and H.J.S. proposed the subdomains for the certainty of evidence assessment. All authors participated in methodological discussions. All authors read and approved the final manuscript.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.05.011.

References

[1] Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;64:383−94.

[2] Guyatt GH, Oxman AD, Schunemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. J Clin Epidemiol 2011;64:380−2.

[3] Schunemann HJ, Best D, Vist G, Oxman AD. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ 2003;169:677−80.

[4] 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.

[5] Balshem H, Helfand M, Schunemann HJ, Oxman AD, Kunz R, Brozek J, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;64:401−6.

[6] Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol 2013;66:158−72.

[7] Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol 2013;66:173−83.

[8] Langendam M, Carrasco-Labra A, Santesso N, Mustafa RA, Brignardello-Petersen R, Ventresca M, et al. Improving GRADE evidence tables part 2: a systematic survey of explanatory notes shows more guidance is needed. J Clin Epidemiol 2016;74:19−27.

[9] Santesso N, Carrasco-Labra A, Langendam M, Brignardello-Petersen R, Mustafa RA, Heus P, et al. Improving GRADE evidence tables part 3: detailed guidance for explanatory footnotes supports creating and understanding GRADE certainty in the evidence judgments. J Clin Epidemiol 2016;74:28−39.

[10] Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, 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.

[11] Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: clinical practice guidelines. BMJ 2016;353:i2089.

[12] Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol 2013;66:719−25.

[13] Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726−35.

[14] Schunemann HJ, Mustafa R, Brozek J, Santesso N, Alonso-Coello P, Guyatt G, et al. GRADE Guidelines: 16. GRADE evidence to decision frameworks for tests in clinical practice and public health. J Clin Epidemiol 2016;76:89−98.

[15] Parmelli E, Amato L, Oxman AD, Alonso-Coello P, Brunetti M, Moberg J, et al. GRADE evidence to decision (EtD) framework for coverage decisions. Int J Technol Assess Health Care 2017;33:176−82.

[16] Burford BJ, Rehfuess E, Schunemann HJ, Akl EA, Waters E, Armstrong R, et al. Assessing evidence in public health: the added value of GRADE. J Public Health (Oxford, England) 2012;34:631−5.

[17] Puhan MA, Schunemann 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.

[18] Morgan RL, Thayer KA, Bero L, Bruce N, Falck-Ytter Y, Ghersi D, et al. GRADE: assessing the quality of evidence in environmental and occupational health. Environ Int 2016;92-93:611−6.

[19] Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106−10.

[20] Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ 2015;350:h870.

[21] Thayer KA, Schunemann HJ. Using GRADE to respond to health questions with different levels of urgency. Environ Int 2016;92-93:585−9.

[22] Schunemann HJ, Hill SR, Kakad M, Vist GE, Bellamy R, Stockman L, et al. Transparent development of the WHO rapid advice guidelines. PLoS Med 2007;4:e119.

[23] Schunemann H, Cuello-Garcia C, Akl EA, Mustafa R, Meerpohl J, Thayer K, et al. GRADE Guidelines: 18. How ROBINS-I and other tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence. J Clin Epidemiol 2018. https://doi.org/10.1016/j.jclinepi.2018.01.012. [Epub ahead of print].

[24] Guyatt GH, Ebrahim S, Alonso-Coello P, Johnston BC, Mathioudakis AG, Briel M, et al. GRADE guidelines 17: assessing

the risk of bias associated with missing participant outcome data in a body of evidence. J Clin Epidemiol 2017;87:14−22.
[25] Kawata AK, Kleinman L, Harding G, Ramachandran S. Evaluation of patient preference and willingness to pay for attributes of maintenance medication for chronic obstructive pulmonary disease (COPD). Patient 2014;7:413−26.
[26] Zhang Y, P A-C, Guyatt G, Yepes-Nuñez JJ, Akl E, Hazlewood G, et al. GRADE guidance for rating the certainty of a body of evidence describing the importance of outcomes or values and preferences: 1. Risk of bias and indirectness. J Clin Epidemiol 2018. https://doi.org/10.1016/j.jclinepi.2018.01.013. [Epub ahead of print].
[27] Joy SM, Little E, Maruthur NM, Purnell TS, Bridges JF. Patient preferences for the treatment of type 2 diabetes: a scoping review. PharmacoEconomics 2013;31:877−92.
[28] Torrance GW. Preferences for health states: a review of measurement methods. Mead Johnson Symp Perinat Dev Med 1982;37−45.
[29] Ryan M, Scott DA, Reeves C, Bate A, Van Teijlingen ER, Russell EM, et al. Eliciting public preferences for healthcare: a systematic review of techniques. Health Technology Assess 2001;5:1−186.
[30] Sepucha K, Ozanne EM. How to define and measure concordance between patients' preferences and medical treatments: a systematic review of approaches and recommendations for standardization. Patient Educ Couns 2010;78:12−23.
[31] 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;64:1294−302.
[32] Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 2008;8:79.
[33] Badia X, Herdman M, Kind P. The influence of ill-health experience on the valuation of health. Pharmacoeconomics 1998;13:687−96.
[34] Brazier J, Rowen D, Karimi M, Peasgood T, Tsuchiya A, Ratcliffe J. Experience-based utility and own health state valuation for a health state classification system: why and how to do it. Eur J Health Econ 2018;19:881−91.
[35] Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010;340:c117.
[36] Umar N, Yamamoto S, Loerbroks A, Terris D. Elicitation and use of patients' preferences in the treatment of psoriasis: a systematic review. Acta Derm Venereol 2012;92:341−6.
[37] 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;64:1283−93.
[38] Schunemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? J Clin Epidemiol 2016;75:6−15.
[39] Jaeschke R, Guyatt GH, Dellinger P, Schunemann H, Levy MM, Kunz R, et al. Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive. BMJ 2008;337:a744.
[40] Guyatt GH, Norris SL, Schulman S, Hirsh J, Eckman MH, Akl EA, et al. Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: antithrombotic therapy and prevention of Thrombosis, 9th ed.: American College of chest Physicians evidence-based clinical practice guidelines. Chest 2012;141:53s−70s.
[41] Zhang Y, Morgan R, Alonso-Coello P, Wiercioch W, Bała M, Jaeschke R, et al. A systematic review on how patients value chronic obstructive pulmonary disease outcomes. Eur Respir J. 2018;52.
[42] 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;64:1277−82.
[43] Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, et al. GRADE guidelines: 9. Rating up the quality of evidence. J Clin Epidemiol 2011;64:1311−6.
[44] Schunemann HJ, Griffith L, Jaeschke R, Goldstein R, Stubbing D, Guyatt GH. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. J Clin Epidemiol 2003;56:1170−6.
[45] Alonso-Coello P, Montori VM, Diaz MG, Devereaux PJ, Mas G, Diez AI, et al. Values and preferences for oral antithrombotic therapy in patients with atrial fibrillation: physician and patient perspectives. Health Expect 2015;18:2318−27.
[46] Gage BF, Cardinalli AB, Albers GW, Owens DK. Cost-effectiveness of warfarin and aspirin for prophylaxis of stroke in patients with nonvalvular atrial fibrillation. JAMA 1995;274:1839−45.
[47] Gage BF, Cardinalli AB, Owens DK. The effect of stroke and stroke prophylaxis with aspirin or warfarin on quality of life. Arch Intern Med 1996;156:1829−36.
[48] Man-Son-Hing M, Laupacis A, O'Connor AM, Coyle D, Berquist R, McAlister F. Patient preference-based treatment thresholds and recommendations: a comparison of decision-analytic modeling with the probability-tradeoff technique. Med Decis Making 2000;20:394−403.
[49] Protheroe J, Fahey T, Montgomery AA, Peters TJ. The impact of patients' preferences on the treatment of atrial fibrillation: observational study of patient based decision analysis. BMJ 2000;320:1380−4.
[50] Slot KB, Berge E. Thrombolytic treatment for stroke: patient preferences for treatment, information, and involvement. J Stroke Cerebrovasc Dis 2009;18:17−22.
[51] Thomson R, Parkin D, Eccles M, Sudlow M, Robinson A. Decision analysis and guidelines for anticoagulant therapy to prevent stroke in patients with atrial fibrillation. Lancet 2000;355:956−62.