Journal of Clinical Epidemiology 111 (2019) 94−104

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

GRADE Guidelines: 19. Assessing the certainty of evidence in the importance of outcomes or values and preferences—Risk of bias and indirectness

Yuan Zhang[a], Pablo Alonso-Coello[a,b,*], Gordon H. Guyatt[a], Juan José 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], Holger J. Schünemann[a,l,*]

[a]*Department of Health Research Methods, Evidence, and Impact & McMaster GRADE Centre, McMaster University, 1280 Main Street West, Hamilton, Ontario L8N 4K1, Canada*

[b]*Centro Cochrane Iberoamericano, Instituto de Investigacion Biomedica (IIB Sant Pau-CIBERESP), Sant Antoni Maria Claret 167, 08025 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, Alberta, Canada*

[e]*Department of Clinical Policy, American College of Physicians, 190 N. Independence Mall West, Philadelphia, PA 19106, USA*

[f]*Center of Innovation for Health Services Research in Primary Care, Durham Veterans Affairs Medical Center and Duke University, Durham, NC 27701, USA*

[g]*Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada*

[h]*Division for Health Services, Norwegian Institute of Public Health, Oslo, Norway*

[i]*Department of Health Management and Health Economics, Institute of Health and Society, University of Oslo, Oslo, Norway*

[j]*Department of Internal Medicine, Division of Nephrology and Hypertension, University of Kansas Medical Center, Kansas City, KS, USA*

[k]*Department of Clinical Epidemiology and Biostatistics, Pontificia Universidad Javeriana, Bogotá, Colombia*

[l]*Department of Medicine, McMaster University, Hamilton, Ontario, Canada*

Accepted 11 January 2018; Published online 13 February 2018

Abstract

Objectives: The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) working group defines patient values and preferences as the relative importance patients place on the main health outcomes. We provide GRADE guidance for assessing the risk of bias and indirectness domains for 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: This is the first of two articles. A body of evidence addressing the importance of outcomes starts at ‘‘high certainty’’; concerns with risk of bias, indirectness, inconsistency, imprecision, and publication bias lead to downgrading to moderate, low, or very low certainty. We propose subdomains of risk of bias as selection of the study population, missing data, the type of measurement instrument, and confounding; we have developed items for each subdomain. The population, intervention, comparison, and outcome elements associated with the evidence determine the degree of indirectness.

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.

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

Conflicts of interest: All authors have completed the ICMJE uniform disclosure form at http://www.icmje.org/conflicts-of-interest/).

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 H.J.S. G.H. 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.

* Corresponding authors. Department of Health Research Methods, Evidence, and Impact and McMaster GRADE Centre, McMaster University Health Sciences Centre, Room 2C16, 1280 Main Street West, Hamilton, Ontario L8N 4K1, Canada. Tel.: +1-905-525-9140x24931; fax: +1-905-522-9507.

E-mail addresses: PAlonso@santpau.cat (P. Alonso-Coello); schuneh@mcmaster.ca (H.J. Schünemann).

https://doi.org/10.1016/j.jclinepi.2018.01.013

Conclusion: This article provides guidance and examples for rating the risk of bias and indirectness for a body of evidence summarizing the importance of outcomes.

Keywords: GRADE; Quality of evidence; Importance of outcomes; Value and preference; Risk of bias; Indirectness

1. Introduction

Decisions in health care require not only evidence about the effects of interventions (eg, the absolute risk reduction or increase for an outcome in a particular population resulting from a specific intervention when compared with an alternative) but also knowledge of the relative importance of the outcomes that interventions prevent or cause (see Box 1 for a hypothetical example).

Incorporating these concepts in health-care decision-making often refers to considerations of values and preferences [1–7]. In the context of decision-making, values and preferences represent the relative importance people place on the outcomes of interest resulting from a decision (eg, about accepting a treatment or undergoing a test) [1–7].

The methods that investigators use to ascertain the relative importance of the outcomes include (a) direct measurement of the utility or value of outcomes, for example, with the standard gamble [8–10], time trade-off [11,12], or rating scales [9,13,14]. Conjoint analysis is another category to elicit utility and indicate outcome importance, which includes discrete choice experiments [15,16], contingent valuation and willingness to pay [17], probability trade-off [18,19], paired comparison; (b) indirect measurement of utility: results from instruments such as the EuroQual-5-dimension (EQ-5D) utility, or Short form-6-dimension (SF-6D) utility, which would transform the measurement results across several domains, that is, pain, mobility, into the utility [20,21]; or (c) other quantitative surveys and questionnaires that provide outcome importance information in a nonutility manner [22,23]. In addition, qualitative studies can provide evidence about the relative importance of outcomes [24,25] (see Appendix 1).

Box 1 A hypothetical example for considering the importance of outcomes

The evidence comparing a new intervention to standard care shows an absolute risk reduction of 10 per 1,000 for a harmful outcome "A" and an absolute risk increase of 10 per 1,000 for a harmful outcome "B".

- If outcomes A and B are judged as equally important (eg, thrombosis and bleeding, respectively), then the balance of benefits and harms does not favor or disfavor the new intervention.
- If outcome A is judged as relatively more important than outcome B (eg, mortality and bleeding, respectively), then the balance of benefits and harms is in favor of the new intervention.

Given health-care decisions will be influenced by both the health effects of interventions, and the relative importance of the outcomes of interest, they both require appropriate methods of certainty assessment. The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group has developed approaches to assess certainty of evidence addressing intervention effects [26,27], test accuracy [28], resources [29], prognosis [30], and qualitative evidence [31]. However, GRADE recognized the increasing need to develop a transparent and structured approach to assess the certainty of evidence for relative importance of the outcomes. Conceptually the GRADE process has required judgments of the certainty in relative importance of the outcomes to draw conclusions in its Evidence to Decision (EtD) tables and frameworks [32–37]. In the last iteration of the GRADE EtD frameworks, the question related to the relative importance of outcomes is "is there important uncertainty about or variability in how much people value the main outcomes?"

Having illustrated the rationale for these considerations, we will describe our terminology (See Box 2) [3,38]. We recognize inconsistent use of the terms in the scientific community. For example, not all scientists agree that a visual analogue scale (VAS) is a utility instrument because it does not require a choice under uncertainty. While acknowledging this fact, we use "outcome importance," which includes but goes beyond the strict definition of "utility." The merit of using "outcome importance" is that it is consistent with the conceptual process of balancing health benefits and harms. In addition, we focus on "relative" importance of outcomes to express that the importance relates to anchors (eg, 0 indicating death, and 1 indicating perfect health) or other outcomes that an intervention causes and which may be balanced against each other to make informed decisions.

The aim of this and the next article related to relative importance of the outcomes is to provide guidance about the GRADE approach for assessing the certainty of a body of evidence dealing with relative importance of the outcomes. In this article, we describe the definitions and methods of this project, and the GRADE approach for rating the domains' risk of bias and indirectness of relative importance of the outcomes. The second article will focus

What is new?

Key findings

- Risk of bias, indirectness, inconsistency, imprecision, and publication bias are domains to be considered to rate down the certainty of evidence.

What this adds to what is known?

- For the risk of bias domain, the subdomains include selection of participants into the study, completeness of data, measurement instrument, and data analysis. Evidence for the relative importance of outcomes may be rated as not serious, serious, or very serious, depending on the contribution of studies with risk of bias concern to the body of evidence. The population, intervention, comparison and outcome elements of the rated evidence and methodological aspects determine the degree of indirectness.

What is the implication and what should change now?

- Authors of knowledge synthesis for the relative importance of outcomes, including utility and values, should consider assessing the certainty they can place on the available research across the body of evidence.

Box 2 Terminology

Terminology	Scope or definition
Outcome	The term outcome includes ''*health state*'' and nonhealth states that are relevant to the alternative treatment under consideration. This includes a broad set of the outcomes directly and indirectly related to health or a disease, an intervention, or nonhealth consequences. Outcomes can be more or less health-related. For example, from mostly health related to least, patients will have their views regarding the importance of the following outcomes: breathlessness, treatment burden of warfarin or insulin injection, ease of reaching a clinic to undergo blood tests and other monitoring.
Relative importance of outcomes	The ***relative importance of outcomes*** is interchangeably used with ***values and preferences, outcome importance,*** or ***outcome valuation but conceptually focuses on the outcomes resulting from an intervention or decision*** [3].
Instrument (for determining the relative importance of outcomes)	This term, when used referring to measure relative importance of outcomes, refers to ''***measurement tool,***'' ''***measurement methods,***'' or ''***measurement instruments***''.
Certainty of evidence	This term is interchangeably used with ''***quality of evidence,***'' ''***strength of evidence,***'' and ''***confidence in estimate***''. ***Certainty of evidence*** has different meanings for systematic reviews and guideline development. For systematic reviews, the definition is the extent of our confidence that the relative importance of the outcomes (and variability) lie in a particular range; for guidelines, the definition is the extent of our confidence that the estimate of the relative importance of the outcomes (and variability) are adequate to support a particular recommendation [38].

on the domains of inconsistency, imprecision, and publication bias and rating up the certainty of evidence. The second article will also clarify what variability of values and preferences or the relative importance of the outcomes means in this context.

2. Methods

This document presents formal guidance by the GRADE working group for rating the relative importance of outcomes. We developed the guidance using an iterative multipronged approach to develop this GRADE guidance. The work was presented at GRADE working group meetings and reviewed by members of the GRADE working group before approval through a vote at a GRADE working group meeting in Rome, Italy, on April 27, 2017. It was then formally approved by the GRADE guidance group.

2.1. Summarizing certainty domains and methods for assessing the certainty of evidence and developing the GRADE approach

Based on a previous systematic survey project [39], we identified systematic reviews addressing the relative importance of the outcomes and qualitatively summarized existing methods used to assess the certainty of a body of evidence and other potential quality indicators, that is, all factors perceived to influence certainty. After discussion, we constructed a list of possible factors and then matched them to the existing GRADE domains. We considered the existing GRADE domains of risk of bias, inconsistency, indirectness, imprecision, and publication bias for downgrading [38]; and large effect sizes, the existence of a dose-response gradient or if residual plausible confounding bias would increase our certainty for upgrading [40]. We planned to record any additional domains that did not fit into existing GRADE domains.

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 [4]. (see Appendix 2 for the full citation of included studies of this systematic review)

Quality assessment							**Estimate of outcome importance (95% CI or other measure of variability)**	**Quality**
Outcome	**Study design/ measurement instrument**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	Other		
Stroke								
Nonfatal severe stroke	Seven cross-sectional studies, 580 participants VAS, SG, TTO	Not serious[a,b,c,d]	No serious inconsistency	No serious indirectness	No serious imprecision	None	0.1–0.39 (range of the point estimates) 0.149, 95% CI: 0.135–0.163	⊕⊕⊕⊕ High
Moderate stroke	Five cross-sectional studies, 339 participants TTO, SG	Not serious	Serious inconsistency[e,f]	No serious indirectness	No serious imprecision	None	0.29–0.77 (range of the point estimates) 0.664, 95% CI: 0.643–0.684	⊕⊕⊕◯ Moderate
Bleeding								
Major (unspecified) GI bleeding	Three cross-sectional studies, 153 participants VAS, TTO, and SG	Not serious[a,c]	No serious inconsistency	No serious indirectness	No serious imprecision	None	0.65–0.84 (range of the point estimates) 0.789, 95% CI: 0.758–0.820	⊕⊕⊕⊕ High
PPS								
Severe PPS	Two cross-sectional studies, 66 participants SG	Not serious[g]	No serious inconsistency	Serious indirectness [h]	Serious imprecision[i]	None	0.93–0.982 (range of the point estimates) 0.973, 95% CI: 0.964–0.982	⊕⊕◯◯ Low
DVT								
DVT and VTE, and bleeding	One cross-sectional study[j], 124 participants Time trade-off	Not serious	No serious inconsistency	No serious indirectness	No serious imprecision	None	If there are a 3% chance of a major bleeding event, and a 2% chance of a recurrent episode of venous thromboembolism in the next 2 yr, the rates of recurrence of DVT without treatment varied from 5%, 10% to 15%, the percentage of participants choosing to stop the VKA treatments are 21%, 23%, and 8%, respectively.	⊕⊕⊕⊕ High
Burden of treatment								
Burden of treatment: warfarin	Seven Cross-sectional studies, 466 participants VAS, SG, TTO	Not serious[a,b,c,d]	No serious inconsistency	No serious indirectness	No serious imprecision	None	0.66–1 (range of estimates across included studies) 0.938, 95% CI: 0.934–0.942	⊕⊕⊕⊕ High

(*Continued*)

Table 1. Continued

Burden of treatment: anticoagulant/ warfarin	One qualitative study, 21 participants Semi-structured interview[k]	Not serious	No serious inconsistency	No serious indirectness	Serious imprecision[l]	None	The majority (specific percentage not reported) of participants had not experienced complications due to warfarin. Many participants reported only minor inconveniences, such as taking a pill every day, regular blood tests, and dietary changes.	⊕⊕⊕○ Moderate

GRADE Working Group grades of evidence.

High quality: We are very confident that the true effect lies close to that of the estimate of the effect.

Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.

Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.

Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect.

Abbreviations: CI, confidence interval; DVT, deep vein thrombosis; GI bleeding, gastrointestinal bleeding; GRADE, Grading of Recommendations Assessment, Development and Evaluation; PPS, postphlebitic syndrome; SG, standard gamble; TTO, time trade-off; VAS, visual analogue scale; VKA, vitamin K antagonists; VTE, venous thromboemblosim.

[a] The representativeness of the studies was impacted by a low response. However, this only impacted a small proportion of the included study population.

[b] In Protheroe 2000, 97 of the 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.

[e] Wide variation across point estimates.

[f] The included study population were patients with atrial fibrillation (Gage 1996), 30 community volunteer (Lenert 1997), three different patient population (patients with a first or second episode of venous thromboembolism, with oral anticoagulants had been started, patients who had experienced an episode of major bleeding during oral anticoagulant treatment, and patients with a postthrombotic syndrome in Locadia 2004), both patients with deep vein thrombosis and without deep vein thrombosis (O'Mera 1994) as well as ischemic stroke survivors and age-matched control subjects (Slot 2009).

[g] One of the studies (Lenert 1997) was judged to be of high risk of bias. However, this study had similar estimates with the other one with low risk of bias.

[h] The certainty of evidence was downgraded for indirectness. The included studies have different population than the patients facing the choice: 30 community volunteer (Lenert 1997), patients with deep vein thrombosis and without deep vein thrombosis (O'Mera 1994).

[i] Small sample size: 66 participants from two studies.

[j] Locadia 2004 is a cross-sectional study interviewing participants with decision analysis.

[k] Dantas 2004 is a qualitative study on the burden of anticoagulant/warfarin treatment.

[l] Only one qualitative study (Dantas 2004) identified to address this phenomenon.

2.2. Application of GRADE approach to examples

We selected a sample of 10 systematic reviews [39] using a maximum variance sampling strategy ensuring that the selection would allow us to illustrate all GRADE domains and address a diversity of health conditions. First, we considered if the existing GRADE domains covered all aspects of certainty for rating the relative importance of the outcomes but did not identify new domains. We adapted the considerations and signaling questions for the assessment from prior GRADE guidance, for example, guidance on prognostic evidence [30]. For every example assessed, we recorded decisions supporting downgrading and developed GRADE evidence profiles (see Table 1 for example) [4,41]. Six investigators (P.A., H.P.H., I.E., J.J.Y.N., Y.Z., and Y.Z.) independently rated the certainty of evidence in pairs using the GRADE domains. We resolved disagreements through discussion or feedback from senior GRADE members, who also evaluated the examples (H.J.S., G.G.). We saved the results as cloud-based documents, and prepared them for comments and feedbacks.

2.3. Consulting for feedback

To ensure a broad perspective, we provided the examples and the guidance to a group of individuals including guideline developers, systematic review and health technology assessment authors, clinical epidemiologists, biostatisticians, clinicians, and researchers with experience in relative importance of outcomes assessment from Canada, the USA, and Europe. We gave the group access to the cloud-based documents and collected feedback from this group through six rounds of online meetings, complemented by emails and in-person meetings, and telephone calls (see Appendix 3 for minutes of the study group meetings). We revised the guidance, documented the adjustments made, and circulated the records for review and comments as part of a GRADE project group. Following each round of feedback, we iteratively improved the preliminary GRADE guidance for assessing the certainty of evidence in the relative importance of outcomes and illustrated the rationale with examples (see Appendix 4 for the guidance in different stages). For example, one major change we made is that we added considerations for subgroup credibility in the inconsistency domain.

As part of the internal review process, we presented the work and guidance at five of the regular meetings of the GRADE working group, and its members had the opportunity to discuss and provide feedback before approval through a vote at a GRADE working group meeting in Rome, Italy, on April 27, 2017. Then, the GRADE guidance group reviewed and approved the document as official guidance before its submission for peer review and publication.

3. Guidance for GRADE domains

We did not identify additional domains for assessing the certainty of the body of evidence describing relative importance of the outcomes, beyond what the GRADE working group had suggested previously: risk of bias, inconsistency, indirectness, imprecision, publication bias, and domains to rate up the evidence [38]. Here, we focus on the detailed guidance for the GRADE domains risk of bias and indirectness.

4. Risk of bias or limitations in the detailed study design or execution

Risk of bias may be a concern at different stages of an investigation into relative importance of the outcomes, including study design, study execution, data analysis, and reporting [42]. Assessing risk of bias for the relative importance of outcomes is similar to that for intervention effects in that it requires first an assessment of the risk of bias for individual studies, followed by an assessment for the body of evidence. However, it differs in several important ways. First, unlike studies on treatment effect, there is no accepted or commonly used tool for assessment of risk of bias for the relative importance of outcomes; or is there a tool that can assess the risk of bias across various study designs [42]. Second, given that the relative importance of an outcome is an estimate that does not represent an effect but is conceptually closer to an estimate of test accuracy or baseline risk. Thus, randomization is not required to protect against confounding bias and balance the known and unknown prognostic factors influencing the outcomes, and certainty of evidence from nonrandomized studies begins as high certainty [28,30].

4.1. *Risk of bias subdomains*

We developed the following subdomains and, for each subdomain, signaling questions, for assessing the risk of bias domain (see Table 2 and Appendix 5 for further detailed guidance):

1. Selection of participants into the study: To what extent does the enrolled study sample reflect the intended population? Improper sample selection will lead to biased estimates of relative importance of outcomes if differing characteristics are associated peoples' relative importance of outcomes.
2. Completeness of data: To what extent are those responding to questions similar to those not responding? High attrition during the follow-up process or low response rate for cross-sectional studies may result in individuals participating who differ systematically in their relative importance of the outcomes from those who do not participate [43,44].
3. Measurement instrument: To what extent a valid instrument has been chosen to elicit the relative importance of the outcomes and has been administered rigorously? This subdomain includes four items: choice of the instrument, administration of the instrument, outcome presentation, and understanding of the instrument by the study population. Low reliability or validity of measurement can result from intrinsic limitations of the measurement instrument or administration error. Moreover, the measurement of outcome importance could be divided into two categories per the nature of the outcome assessed: measurement of respondents' outcome, which they are experiencing or have experienced, or measurement of described (usually hypothetical) outcomes, which they may or may not experience in the future. This involves the judgment of indirectness, which we will describe later. However, in the latter case, another consideration independent of indirectness relates to how valid the presentation of the outcome is.

Table 2. Risk of bias subdomains and signaling questions

Subdomain	Signaling questions
Selection of participants into the study	Was an appropriate study sample selected from the sampling frame?
Completeness of data	Was the attrition sufficiently low to minimize the risk of bias?
Measurement instrument	Was the instrument used for eliciting relative importance of outcomes valid and reliable? Was the instrument administered in the intended way? Was a valid representation of the outcome (health state) utilized? Did the researchers check the understanding of the instrument?
Data analysis	Were the results analyzed appropriately to avoid influence of bias and confounding?

4. Data analysis: To what extent the estimate is distorted by inappropriate data analysis? Was adjustment, stratification in the analysis and model selection, if any, appropriate to avoid distorted results from confounding?

Based on answers to the above signaling questions, each study, depending on the likelihood of bias and the magnitude of its impact on the estimates, is classified as low, moderate, serious, and critical risk of bias for each subdomain (see Box 3). The classification of risk of bias of individual studies (across these subdomains within a study, see Box 4) is helpful for describing individual studies and is required for an assessment across studies,that is, the body of evidence.

4.2. Summary of risk of bias

The decision for risk of bias at the body of evidence level requires inspecting the overall pattern of results across subdomains and studies, the relative weight or contribution of studies with risk of bias concern, and whether the risk of bias is likely to influence the overall results. The assessment at the body of evidence level is labeled not serious, serious, or very serious.

We encourage raters to attempt making a judgment based on the information available (either in the study report or after obtaining additional information from authors), and including inferences about what is not stated, but is most likely.

Consistent with the GRADE approach for other types of evidence, the risk of bias assessment is conducted for each outcome [45]. If most information is from studies at low risk of bias for all subdomains, the overall judgment of risk of bias should be "low risk of bias," and in the GRADE certainty of evidence, raters would not downgrade. However, as the contribution of studies with risk of bias concerns to the body of evidence increases, raters downgrade the certainty of evidence by one or more levels due to risk of bias [46]. Risk of bias assessment on any subdomain is a continuum, and reviewers must bear this in mind when making their overall judgments. If necessary, raters could conduct sensitivity analysis that evaluates whether or not risk of bias in individual studies is likely to influence the overall results, across subdomains and studies.

Box 3 Judgment of risk of bias for risk of bias subdomains

Response option	Interpretation
Low risk of bias	The estimate for this relative importance of outcomes study is unlikely to be biased with regard to this subdomain.
Moderate risk of bias	The estimate for this relative importance of outcomes study is likely to be biased with regard to this subdomain, but the influence of the bias is limited.
Serious risk of bias	The estimate for this relative importance of outcomes study is probably biased with regard to this subdomain, and the influence of the bias is substantial.
Critical risk of bias	The estimate for this relative importance of outcomes study is certain to be distorted with regard to this subdomain and the estimate is not trustworthy.

Box 4 Overall risk of bias for a study

Response option	Interpretation
Low risk of bias	The study is classified as with low risk of bias across subdomains.
Moderate risk of bias	The study is classified as low or moderate risk of bias across subdomains.
Serious risk of bias	The study is classified as serious risk of bias for at least one subdomain but not classified as critical risk of bias for any subdomain.
Critical risk of bias	The study is classified as critical risk of bias for at least one subdomain.

Example: One systematic review summarized the utility that patients with severe nonfatal strokes placed on their own health [4]. Two of the seven included studies reported a low response rate, and 17% of participants in a third study had difficulties to understand the instrument; these three studies contributed approximately 35% of all participants who provided information for the estimates. However, because no other concern was raised for other risk of bias subdomains, and the results from studies with risk of bias concerns were similar to those at low risk of bias, we did not downgrade for risk of bias [4]. In another systematic review to assess the patient preferences for type 2 diabetes treatment-related outcomes, of all 61 included studies, only six showed that the respondents were similar to nonrespondents [47]. Thus, we downgraded the certainty of evidence due to risk of bias resulting from selection of participants into the studies.

5. Indirectness

Indirectness can be a reason to rate down the certainty of evidence in relative importance of the outcomes [48]. The assessment of indirectness for relative importance of the outcomes has its specific features. First, studies usually do not directly compare the intervention options; rather, the focus is on outcomes. Second, surrogate outcomes or outcomes that are not patient-important are a source of indirectness for treatment questions—this may not be the case for the evidence of relative importance of the outcomes. In importance studies, the outcomes could be indirect because the outcomes may not be representative. Thus, if we are interested in the importance of a surrogate outcome from the patients' perspective, being a "surrogate" does not justify rating down the certainty of evidence. In addition, there is no indirect comparison in the relative importance of outcomes evidence. Finally, the methods to elicit the relative importance of outcomes could be a source of indirectness. Here, we provide the rationale and examples for these considerations, which we organize into two categories: indirectness due to population, intervention, comparison, and outcome (PICO) elements and indirectness due to methodological elements (Table 3).

5.1. PICO elements

For a systematic review addressing the relative importance of outcomes, we could define the research question as "what is the relative importance that patients place on the outcomes when they make a decision related to…", for which we still need clearly defined PICO elements. PICO elements could be sources of indirectness when the body of evidence does not represent the PICO elements of interest (see Appendix 6).

If the outcomes considered in the available studies are not representative of the outcomes of interest, the confidence placed on the evidence is necessarily lower. Whether the intervention and comparison options are a source of indirectness depends on to what extent the outcome considered is different when it is incurred by one intervention versus another. Interventions may differ in many aspects—surgical skills or approaches, or drug dosages, durations, or routes of administration route, but we are only concerned if the differences in the interventions would probably cause different outcomes.

We include the intervention and comparison options, the I and C in PICO for comparison, as a relevant consideration when assessing indirectness (and also in the inconsistency domain) for the following reasons. First, the difference in comparisons may signal potential differences in the type of outcomes. Second, empirical evidence suggests that respondents process the same outcome differently if they are aware that the same outcome is the consequence of different interventions [49].

Example of indirectness due to PICO elements: A systematic review summarized the relative importance patients placed on health states associated with benign prostatic hyperplasia: the assessment of symptom improvement, decreased prostate size, risks of acute urinary retention (AUR), and surgery [50]. It suggested that men would wait longer for symptom improvement in exchange for decreased prostate size (13 months) than they would in exchange for an absolute 1% decrease in the risks of AUR (2 months) and surgery (8 months). However, this valuation was based on one study included in the systematic review, and in this study, 208 men aged ≥40 years from the general population were included. We consider the optimal study population in this case as an aging male population who are at risk of benign prostatic hyperplasia. Thus, the study did not enroll an optimal study population because men (male ≥ 40 years) were generally younger than the population who are typically facing the decision. We rated the certainty of evidence down for indirectness of the population because the trade-off and valuation of outcomes involve AUR and surgery, which are usually not the decision most men aged 40 years and older from general population would make [50]. Meanwhile, although not optimal, aging males from the general population are at risk of prostatic hyperplasia, and the presented considerations are not completely irrelevant for them. For this PICO, we identified no other concern for indirectness. As this example demonstrates, the merit of GRADE approach is not to eliminate disagreement but rather to provide a transparent and explicit assessment process.

Table 3. Signaling questions for indirectness

Sources of indirectness	Signaling questions
Indirectness due to PICO elements	Was the population studied matching the population of interest? Were the outcomes matching the outcomes of interest? Were the options studied matching the alternative options of interest?
Indirectness due to methodological elements	Were the participants answering questions directly valuing the relative importance of outcomes? • Were direct methodologies for outcome utilities rather than indirect methodologies used? • Was the utility directly estimated from an instrument to elicit utility rather than mapped from instruments whose purpose are not eliciting utility?

5.2. Methodological aspects

The methods used to elicit relative importance of outcomes may also represent a source of indirectness (see Appendix 7). This consideration is applicable whenever investigators have used an indirect measurement technique (ie, a multiattribute utility index) to measure the utility of outcomes (utilities from EQ-5D, SF-6D, quality of well-being, or health utility index) or when a mapping algorithm was used to estimate generic utilities based on the estimates from other measurements (ie, estimating EQ-5D utility from St George Respiratory Questionnaire). This may be based on a linkage or mathematical transformation functions that are used to calculate relative importance of the outcomes based on tools, such as quality of life instruments [51].

When one asks patients to rate the value they place on health states, one can ask the question directly how much value they place on their own health state, or a clinical scenario, using the standard gamble, time trade-off, and VAS. Multiattribute utility measurement instruments (eg, EQ-5D utility, SF-6D utility) have used such direct measurement instrument, together with measurements on health domains (eg, pain, mobility, and so forth) to develop scoring systems for health state ratings, which is the algorithm to help transform measurements on health domains into utility. Users of multiattribute utility measurement instruments then ask respondents only to describe their own health state with the health domains. Thus, respondents are not providing their own evaluation of importance but simply providing information about their experience. To obtain utility of their health states, researchers need an algorithm estimated based on another population. These values then come from someone else, and depending on the type of population, the rating of utility may be indirect.

Essentially the same situation exists when researchers convert disease-specific quality of life scores (eg, St George Respiratory Questionnaire) into generic utilities. In this case, indirect utilities are not estimated, but predicted from research results obtained using an instrument whose purpose was to assess the magnitude of disability, not to estimate the target measurement. Again, the values come from someone else and are indirect.

However, depending on the perspective taken in the health care decision-making process, either in a healthcare policy decision-making scenario, a clinical guideline development project, or a decision for an individual patient, users may or may not judge the indirectness severe enough to rate down. If one accepts that the population completing a multiattribute utility instrument has the same relative importance of outcomes as the individuals who participated in the scenario rating that led to the weighting algorithm in the first place, then one might infer that ratings are those that would be provided by a direct assessment of relative importance of outcomes. Making this assumption, one would not rate down due to indirectness.

5.3. Different strategies for systematic review authors and guideline panellists

In most cases, systematic review authors would only include studies in which the population, intervention and comparisons, and outcomes meet the eligibility criteria, thus assuring directness [48]. However, in some situations, systematic review authors may include indirect evidence and rate down for indirectness concerning their population and outcome of primary interest. In contrast to systematic reviews, use of indirect evidence is very common in the setting of clinical practice guidelines.

These different purposes of using and considering evidence could lead to the different indirectness judgment for the same body of evidence. As previously discussed, for a systematic review addressing the utility of bleeding, a major bleeding that happens after taking aspirin is no more indirect compared with a major bleeding after taking warfarin. In contrast, in guideline development, whether the participants were valuing the importance of bleeding after taking warfarin or after taking aspirin may matter if the severity or type of bleeding differs.

6. Summary

This article describes the use of GRADE to assess the certainty of evidence for the relative importance of outcomes when considering risk of bias and indirectness. When assessing certainty of evidence for the relative importance of outcomes, evidence starts at ‘‘high’’ for all study designs, with rating down if risk of bias or indirectness is a serious concern. Users rate down by one or more levels depending on the specific considerations for the two domains.

Risk of bias assessment in this context presents challenges. We have proposed a guiding set of questions to consider risk of bias issues; the reliability or validity of our suggested approaches remains unaddressed, but no well-validated instruments exist so far. Pending this work, using the signaling questions and examples we have provided will help make judgments regarding risk of bias transparent.

In the next article, we will discuss the application of the other GRADE domains (imprecision, inconsistency, publication bias, and upgrading domains) in the assessment of certainty of relative importance of outcomes evidence.

Acknowledgments

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

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

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2018.01.013.

References

[1] Schunemann HJ, Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 10. Integrating values and consumer involvement. Health Res Policy Syst 2006;4:22.

[2] Murad MH, Montori VM, Guyatt GH. Incorporating patient preferences in evidence-based medicine. JAMA 2008;300:2483. author reply-4.

[3] Schunemann HJ, Wiercioch W, Etxeandia I, Falavigna M, Santesso N, Mustafa R, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. CMAJ 2014;186:E123–42.

[4] MacLean S, Mulla S, Akl EA, Jankowski M, Vandvik PO, Ebrahim S, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: antithrombotic therapy and prevention of thrombosis, 9th ed.: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;141:e1S–23S.

[5] 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.

[6] Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. JAMA 2008;300(4):436–8.

[7] van der Weijden T, Legare F, Boivin A, Burgers JS, van Veenendaal H, Stiggelbout AM, et al. How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. Implement Sci 2010;5:10.

[8] Gafni A. The standard gamble method: what is being measured and how it is interpreted. Health Serv Res 1994;29:207–24.

[9] Torrance GW. Measurement of health state utilities for economic appraisal. J Health Econ 1986;5(1):1–30.

[10] Torrance GW. Utility measurement in healthcare: the things I never got to. PharmacoEconomics 2006;24(11):1069–78.

[11] Churchill DN, Torrance GW, Taylor DW, Barnes CC, Ludwin D, Shimizu A, et al. Measurement of quality of life in end-stage renal disease: the time trade-off approach. Clin Invest Med 1987;10(1):14–20.

[12] Dolan P, Gudex C, Kind P, Williams A. The time trade-off method: results from a general population study. Health Econ 1996;5(2):141–54.

[13] Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? Med Decis Making 2001;21:329–34.

[14] Morimoto T, Fukui T. Utilities measured by rating scale, time trade-off, and standard gamble: review and reference for health care professionals. J Epidemiol 2002;12(2):160–78.

[15] Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. Appl Health Econ Health Policy 2003;2(1):55–64.

[16] Ryan M. Discrete choice experiments in health care. BMJ 2004;328:360–1.

[17] Stevens TH, Belkner R, Dennis D, Kittredge D, Willis C. Comparison of contingent valuation and conjoint analysis in ecosystem management. Ecol Econ 2000;32(1):63–74.

[18] 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 2014;18:2318–27.

[19] Devereaux PJ, Anderson DR, Gardner MJ, Putnam W, Flowerdew GJ, Brownell BF, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. BMJ 2001;323:1218–22.

[20] Craig BM, Busschbach JJ, Salomon JA. Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. Med Care 2009;47:634–41.

[21] Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001;33(5):337–43.

[22] 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.

[23] King M, Nazareth I, Lampe F, Bower P, Chandler M, Morou M, et al. Conceptual framework and systematic review of the effects of participants' and professionals' preferences in randomised controlled trials. Health Technol Assess 2005;9:1–186.

[24] Cronin M, Meaney S, Jepson NJ, Allen PF. A qualitative study of trends in patient preferences for the management of the partially dentate state. Gerodontology 2009;26(2):137–42.

[25] DeJean D, Giacomini M, Vanstone M, Brundisini F. Patient experiences of depression and anxiety with chronic disease: a systematic review and qualitative meta-synthesis. Ont Health Technol Assess Ser 2013;13(16):1–33.

[26] 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.

[27] 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;336:924–6.

[28] 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.

[29] Brunetti M, Shemilt I, Pregno S, Vale L, Oxman AD, Lord J, et al. GRADE guidelines: 10. Considering resource use and rating the quality of economic evidence. J Clin Epidemiol 2013;66:140–50.

[30] 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.

[31] Lewin S, Glenton C, Munthe-Kaas H, Carlsen B, Colvin CJ, Gulmezoglu M, 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(10).

[32] Akl EA, Grant BJ, Guyatt GH, Montori VM, Schunemann HJ. A decision aid for COPD patients considering inhaled steroid therapy: development and before and after pilot testing. BMC Med Inform Decis Mak 2007;7:12.

[33] Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ 2008;336:1049–51.

[34] 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.

[35] 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.

[36] 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(5):e119.

[37] Kelson M, Akl EA, Bastian H, Cluzeau F, Curtis JR, Guyatt G, et al. Integrating values and consumer involvement in guidelines with the patient at the center: article 8 in Integrating and coordinating efforts in COPD guideline development. An official ATS/ERS workshop report. Proc Am Thorac Soc 2012;9(5):262–8.

[38] 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.

[39] Yepes-Nuñez JJ, Zhang Y, Xie F, Alonso-Coello P, Selva A, Schünemann H, et al. Forty-two systematic reviews generated 23 items for assessing the risk of bias in values and preferences' studies. J Clin Epidemiol 2017;85:21−31.

[40] 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.

[41] 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.

[42] Karanicolas PJ, Montori VM, Devereaux PJ, Schunemann H, Guyatt GH. A new ''mechanistic-practical'' framework for designing and interpreting randomized trials. J Clin Epidemiol 2009;62:479−84.

[43] Levin KA. Study design III: cross-sectional studies. Evid Based Dent 2006;7(1):24−5.

[44] Fincham JE. Response rates and responsiveness for surveys, standards, and the Journal. Am J Pharm Educ 2008;72:43.

[45] Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;64:395−400.

[46] 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;64:407−15.

[47] 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(10):877−92.

[48] 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;64:1303−10.

[49] Holbrook A, Labiris R, Goldsmith CH, Ota K, Harb S, Sebaldt RJ. Influence of decision aids on patient preferences for anticoagulant therapy: a randomized trial. CMAJ 2007;176(11):1583−7.

[50] Emberton M. Medical treatment of benign prostatic hyperplasia: physician and patient preferences and satisfaction. Int J Clin Pract 2010;64:1425−35.

[51] Brazier JE, Yang Y, Tsuchiya A, Rowen DL. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. Eur J Health Econ 2010;11(2):215−25.