GRADE UPDATE OF PAPERS

The GRADE Working Group clarifies the construct of certainty of evidence

Monica Hultcrantz\textsuperscript{a,b,*}, David Rind\textsuperscript{c,d}, Elie A. Akl\textsuperscript{e,f}, Shaun Treweek\textsuperscript{g}, Reem A. Mustafa\textsuperscript{e,h}, Alfonso Iorio\textsuperscript{c,d}, Brian S. Alper\textsuperscript{i,j,k}, Joerg J. Meerpohl\textsuperscript{l,m}, M Hassan Murad\textsuperscript{n}, Mohammed T. Ansari\textsuperscript{o}, Srinivasa Vittal Katikireddi\textsuperscript{p}, Pernilla Östlund\textsuperscript{q,r}, Sofia Tranæus\textsuperscript{a,q,r}, Robin Christensen\textsuperscript{q}, Gerald Gartlehner\textsuperscript{t,u}, Jan Brozek\textsuperscript{c,d}, Ariel Izcoivich\textsuperscript{q}, Holger Schünemann\textsuperscript{c,e,i}, Gordon Guyatt\textsuperscript{c,e,i}

\textsuperscript{a}Swedish Agency for Health Technology Assessment and Assessment of Social Services (SBU), S:t Eriksgatan 117, SE-102 33, Stockholm, Sweden
\textsuperscript{b}Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Tomtebodav 18 A, SE-171 77, Stockholm, Sweden
\textsuperscript{c}Institute for Clinical and Economic Review, Two Liberty Square, 9th Floor, Boston, MA 02109, USA
\textsuperscript{d}Harvard Medical School Beth Israel Deaconess Medical Center, 330 Brookline Avenue Boston, MA 02215, USA
\textsuperscript{e}Department of Health Research Methods, Evidence, and Impact, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada
\textsuperscript{f}Department of Medicine, American University of Beirut Medical Center, P.O. Box 11-0236, Riad-El-Solh, Beirut 1107 2020, Lebanon
\textsuperscript{g}Health Services Research Unit, University of Aberdeen, 3rd Floor, Health Sciences Building, Foresterhill, Aberdeen, AB25 2ZD, UK
\textsuperscript{h}Department of Medicine, Division of Nephrology and Hypertension, University of Kansas Medical Center, 3901 Rainbow Blvd, MS3002, Kansas City, KS 66160, USA
\textsuperscript{i}Department of Medicine, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada
\textsuperscript{j}EBSCO Health, EBSCO Publishing, 10 Estes Street, Ipswich, MA, 01938 USA
\textsuperscript{k}Department of Family and Community Medicine, School of Medicine, University of Missouri-Columbia, One Hospital Drive, Columbia, MO 65212, USA
\textsuperscript{l}Cochrane Germany, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Breisacher Strasse 153, 79110 Freiburg, Germany
\textsuperscript{m}Centre de Recherche Épidémiologie et Statistique Sorbonne Paris Cité – U1153, Inserm / Université Paris Descartes; Cochrane France; Hôpital Hôtel-Dieu, 1 place du Parvis Notre Dame, 75181 Paris Cedex 04, France
\textsuperscript{n}Mayo Clinic Evidence-based Practice Center, Mayo Clinic, 200 1st Street SW, Rochester, MN 55905, USA
\textsuperscript{o}School of Epidemiology, Public Health and Preventive Medicine, Faculty of Medicine, University of Ottawa, 600 Peter Morand Crescent, Ottawa, Ontario K1G 5Z3, Canada
\textsuperscript{p}MRC/CSO Social & Public Health Sciences Unit, University of Glasgow, 200 Renfield Street, Glasgow, G2 3QB, UK
\textsuperscript{q}Health Technology Assessment - Odontology (HTA-O), Faculty of Odontology, Karl Gustafs Väg 34, SE-214 21, Malmö, Sweden
\textsuperscript{r}Division of Cardiology and Endodontology, Department of Medicine, Karolinska Institutet, Alfred Nobels Allé 8, SE-141 52, Huddinge, Sweden
\textsuperscript{s}Musculoskeletal Statistics Unit, The Parker Institute, Bispebjerg and Frederiksborg Hospital, Nordre Fasanvej 57, DK-2000 Copenhagen F, Denmark
\textsuperscript{t}Danube University, Department for Evidence-based Medicine and Clinical Epidemiology, Dr. Karl Dorreklasse 30, 3500 Krems, Austria
\textsuperscript{u}RTI International, Research Triangle Park, 3040 East Cornallid Rd, Research Triangle Park 27709, NC, USA
\textsuperscript{v}Internal Medicine Service, German Hospital, Pueyrredón 1640, Buenos Aires C1118AA, Argentina

* Corresponding author. Tel.: +46 (0)8-412 32 73 ; fax: +46 (0)8-411 32 60.
E-mail address: monica.hultcrantz@sbu.se (M. Hultcrantz).

Funding: The Health Services Research Unit, University of Aberdeen, receives core funding from the Chief Scientist Office of the Scottish Government Health Directorates. The Parker Institute, Bispebjerg and Frederiksborg Hospital, is supported by a core grant from the Oak Foundation (OCAY-13-309). S.V.K. is funded by an NRS Scottish Senior Clinical Fellowship (SCAF/15/02), the Medical Research Council (MC_UU_12017/13 and MC_UU_12017/15), and Chief Scientist’s Office (SPHSU13 and SPHSU15).

Abstract

Objective: To clarify the grading of recommendations assessment, development and evaluation (GRADE) definition of certainty of evidence and suggest possible approaches to rating certainty of the evidence for systematic reviews, health technology assessments, and guidelines.

Study Design and Setting: This work was carried out by a project group within the GRADE Working Group, through brainstorming and iterative refinement of ideas, using input from workshops, presentations, and discussions at GRADE Working Group meetings to produce this document, which constitutes official GRADE guidance.

Results: Certainty of evidence is best considered as the certainty that a true effect lies on one side of a specified threshold or within a chosen range. We define possible approaches for choosing threshold or range. For guidelines, what we call a fully contextualized approach

Accepted 2 May 2017; Published online 18 May 2017
requires simultaneously considering all critical outcomes and their relative value. Less-contextualized approaches, more appropriate for systematic reviews and health technology assessments, include using specified ranges of magnitude of effect, for example, ranges of what we might consider no effect, trivial, small, moderate, or large effects.

**Conclusion:** It is desirable for systematic review authors, guideline panelists, and health technology assessors to specify the threshold or ranges they are using when rating the certainty in evidence. © 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

**Keywords:** GRADE; Certainty of evidence; Thresholds; Guidelines; Systematic reviews; Health technology assessment

---

### 1. Introduction

The grading of recommendations assessment, development and evaluation (GRADE) Working Group has designed a widely adopted structure for the development of clinical practice and public health guidelines [1]. Formally assessing the trustworthiness of the available evidence represents a key component of the GRADE approach. GRADE offered a formal definition of certainty of evidence: “the extent of our confidence that the estimates of the effect are correct or are adequate to support a particular decision or recommendation” (Box 1) [1]. This definition suggests we are rating confidence or certainty in point estimates of effect. GRADE did not, however, present a coherent conceptual basis for rating certainty in such estimates.

The aim of this article is to present an alternative, and we believe more satisfactory, conceptualization. This alternative is grounded in the realization that, when deciding whether evidence regarding intervention effects is adequate to support a recommendation, we are not assessing our confidence in point estimates of effects, but rather our confidence in where effects lie relative to particular thresholds [2].

In the first part of our discussion, we make evident that thresholds depend on the health care context and therefore can vary. In the second part of our discussion, we examine how authors can formally set thresholds, or ranges they are using when rating the certainty in evidence.

In the context of making recommendations, the quality of evidence reflect the extent of our confidence that the estimates of the effect are correct. In the context of making recommendations, the quality ratings reflect the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation.

With regard to the necessity for thresholds and the possibility they may vary, consider a societal choice to invest in the platelet inhibitor ticagrelor for patients after myocardial infarction. In a hundred typical patients, the drug may prevent one death over the course of a year, with limited adverse effects. In high-income countries, the threshold for implementing ticagrelor may be less than a 1% mortality reduction, and policy makers may decide on widespread implementation. In this context, we may have high certainty that the benefit exceeds our threshold for supporting the recommendation. In low-income countries, the opportunity cost of offering ticagrelor is likely to be prohibitive, and the threshold for implementation may be much higher (a 10% mortality reduction, or perhaps even larger). In this context, we may have low certainty that the benefit exceeds our threshold for supporting the recommendation and may even have high certainty that the benefit fails to exceed a threshold for supporting the recommendation and could warrant supporting a recommendation against.

Thus, given the same evidence regarding a particular outcome, in the context of a guideline, the certainty of the evidence can vary depending on the context, and the health care question being asked. This is often surprising to those who first encounter the concept. Therefore, in the first part of the discussion in the following, we will present another example illustrating this crucial concept.

#### 1.1. Definitions

GRADE initially referred to “quality of evidence”; subsequently “confidence in the estimates” replaced “quality of evidence”; most recently “certainty of evidence” has often become the preferred term. These words all refer to the same concept, and we will use “certainty of evidence” throughout this article.

As discussed previously, certainty of evidence defined as adequacy to support a particular decision or recommendation varies with the health care context. We will refer to situations where the full health care question/context is made clear as “fully contextualized.” Such fully contextualized ratings are typically made in the setting of clinical practice guidelines. We will discuss the distinctions between assessments made that are fully contextualized, partly contextualized, and noncontextualized—the latter two typically made in context of systematic reviews and

---

**Box 1** GRADE’s adopted definition of certainty of the evidence [1]. Note that “quality of evidence” refers to the same concept as “certainty of evidence” (see paragraph on definitions).

In the context of a systematic review, the ratings of the quality of evidence reflect the extent of our confidence that the estimates of the effect are correct. In the context of making recommendations, the quality ratings reflect the extent of our confidence that the estimates of an effect are adequate to support a particular decision or recommendation.
**What is new?**

**Key findings**
- The grading of recommendations assessment, development and evaluation (GRADE) Working Group clarifies that when rating certainty of the evidence for an individual outcome, we are rating how certain we are that the true effect lies within a particular range or on one side of a threshold.
- We illustrate approaches for setting thresholds or ranges with different degrees of contextualization that can be used for a systematic review, health technology assessment, or guideline.

**What this adds to what was known?**
- GRADE’s published definition of certainty of evidence leaves some ambiguity: does it refer to confidence in point estimates, or confidence in a particular range in which the true effect may lie? Our presentation provides a clear and coherent answer to this question.

**What is the implication and what should change now?**
- It is desirable for systematic review authors, guideline panelists, and health technology assessors to specify the threshold or ranges they are using to rate their certainty of the evidence for an individual outcome. When choosing an approach for setting threshold or range, the authors should consider the setting and what would be of most use for the target audience.

If there are no serious concerns about risk of bias, inconsistency, indirectness, or publication bias, the CI will represent a reasonable estimate of a certainty range, the range of reasonably believable effects of the intervention; if there were such concerns, the certainty range (i.e., the range in which we anticipate the true effect may lie, after considering not only precision but risk of bias, inconsistency, indirectness, and publication bias) would be wider and/or shifted compared to the 95% CI, although its exact distribution would be difficult to ascertain [4].

This thought exercise occurs in the setting of a fully contextualized evidence certainty rating (typically, a guideline setting). We ask readers to first assume that the intervention is a drug with no serious adverse effects, minimal inconvenience, and modest cost that is equitable, feasible, and acceptable to administer. Under these circumstances, even a small beneficial health effect would warrant a recommendation for the intervention because, overall, the desirable consequences would outweigh the undesirable consequences. For instance, given the considerations about all the possible downsides or harms, we may recommend the intervention if it reduced the incidence of stroke by as little as 0.5% (vertical green line in Fig. 1). Here, we set a threshold, 0.5%, that defines our willingness to recommend the intervention.

The entire CI (0.6–2.0%) around the effect on stroke reduction lies to the left of the clinical decision threshold of 0.5% and therefore excludes a benefit smaller than the threshold. We can—as we point out in the article—therefore conclude that the precision of the estimate is sufficient to support a recommendation: we are confident that the true effect lies above our threshold and there is no reason to rate down certainty as a result of imprecision. Assuming there are no serious concerns with risk of bias, inconsistency, indirectness, or publication bias, our certainty that the true effect lies above our threshold will be high, and any ensuing recommendation is likely to be strong.

Consider now a modification of this hypothetical scenario in which the same treatment is associated with more serious

---

2. Rating certainty: the concept

2.1. We are not rating certainty in point estimates, but rather certainty that the true effect lies in a particular range: an illustration from previous GRADE writings

An illustration of rating our certainty that the effect lies above a particular threshold in GRADE is in the 6th article of the JCE series that deals with imprecision [2]. In that article, we present a hypothetical systematic review of randomized control trials of an intervention to prevent major strokes that yields a pooled estimate of absolute reduction in strokes of 1.3%, with a 95% confidence interval (CI) of 0.6–2.0% (Fig. 1).

---

Fig. 1. Rating certainty that the true effect lies in a particular range: an illustration from previous GRADE writings [2].
harm, such as a 0.5% absolute increase in myocardial infarction. Under these circumstances, we may be reluctant to recommend treatment unless the absolute stroke reduction was larger, for instance at least 1% (red line in Fig. 1). Because the point estimate of 1.3% meets the threshold criterion, a recommendation in favor of treatment would still be appropriate, but the imprecision-generated uncertainty will result in only moderate certainty that the effect is above the threshold. Not only does it lead to lowering our certainty in the evidence, in this situation we are also likely to make a weak recommendation.

In this example, the certainty in effect size for the effect of stroke reduction does not change, but as our threshold for treatment changes (as the magnitude of undesirable health effects increases), the certainty that desirable consequences outweigh undesirable consequences decreases.

The logic of the previous example applies to all fully contextualized graded recommendations: given the undesirable consequences of an intervention, how certain are we that the health benefits lie above a threshold that makes it worthwhile to administer that intervention? Consider patients with atrial fibrillation who are deciding whether to use anticoagulant therapy to lower their risk for stroke. Such patients must weigh the anticipated stroke reduction against the undesirable health effects of anticoagulation, including the risk of bleeding. Thus, patients must ask themselves: how small a stroke reduction am I ready to accept given the bleeding risk associated with anticoagulation, and still use the drug? Depending on the bleeding risk, and patients' values, the answer may be an absolute reduction in stroke risk of 1%, 2%, 3%, or more. Furthermore, patients need to consider other undesirable health effects and consequences of anticoagulation: the burden of medication use, including lifestyle limitations and, if using warfarin, the need for monitoring.

Note, the question patients are asking relates not to the point estimate but rather to the possible range in which the true effect lies. For instance, let us say a patient has chosen a threshold of 2%. For the decision to use or not use anticoagulation, it is immaterial whether the true effect is a 2.1% reduction in stroke or a 3% or larger reduction, as long as it is more than the 2% threshold. Thus, we should ascertain our certainty that the true reduction is ≥2%, not our certainty that the point estimate represents the true effect. Although the point estimate might be above the relevant threshold, if the CI crosses the threshold, we would be less certain that the true effect is above the threshold (and thus rate down for imprecision).

We have framed the example as the magnitude of benefit required. It could as easily be framed as the acceptable magnitude of undesirable consequences [5]. For instance, given a particular reduction in stroke, how great could the magnitude of increased bleeding be before patients would choose to forego anticoagulants.

We have discussed this logic in the context of individual patient decision making. The process is, however, identical for a guideline panel. The panel members must ask themselves the extent to which they are certain that the desirable and undesirable consequences lie in a range that would clearly mandate a recommendation for or against a particular management strategy. They are not therefore—as one might infer from the definition of certainty in Box 1—rating their certainty in the point estimates of effect.

### 2.2. Considering uncertainty in both benefits and harms presents serious challenges to setting thresholds

To this point, we have focused on uncertainty in the benefits and not uncertainty in the harms. Simultaneously considering uncertainty in both benefits and harms raises additional challenges. Moreover, and perhaps even more challenging, we have focused on a single benefit outcome. What if, as is often the case, we have more than one benefit and more than one harm outcome associated with the intervention? The cognitive challenge of simultaneously considering quantitative thresholds when there are multiple desirable and undesirable outcomes is formidable—indeed, possibly beyond the capacities for all but a very few individuals. This is analogous to considering our certainty in total net benefit, as might be generated by a decision analysis model. The challenges do not, however, bear on the essential point that when we rate certainty in evidence, the process has to do with our certainty that true effects lie in a particular range or on one side of a particular threshold.

### 3. Rating certainty: the options

#### 3.1. Implementing the range/threshold approach to rating certainty of evidence—contextualization

The atrial fibrillation example we have used represents a real (albeit simplified) clinical decision with the simultaneous consideration of all critical desirable and undesirable consequences of treatment. Setting the threshold for use of anticoagulants, and thus the rating of certainty, requires a judgment regarding the relative desirability of avoiding stroke, bleeding, and the burden associated with anticoagulation. We refer to patients’ judgments of relative desirability as value and preference judgments. A guideline panel will consider typical values and preferences for the patient group of interest.

In a clinical context, all outcomes associated with a given decision and the associated value and preference judgments are considered simultaneously, and so, a decision about the tradeoff is “fully contextualized.” Clinical practice guideline panels should always be considering such fully contextualized settings, as may systematic reviews that are undertaken specifically to inform a guideline panel (Table 1).

For both health technology assessments and systematic reviews, there is often a certain degree of
contextualization of the results (e.g., by the choice of outcomes presented, the consideration of indirectness, and some notion of how the target audience values the outcomes). Thus, systematic review and health technology assessment authors should be explicit about the context they have in mind, and whether or not issues of feasibility and equity are influencing their judgments (usually, they will not). The specification of a complete set of values and preferences that allow the tradeoff between desirable and undesirable outcomes is not, however, in the authors’ purview. Their consideration of certainty in evidence is not, therefore, fully contextualized.

GRADE writings have recognized this key distinction between practice guidelines and systematic reviews and therefore offered two definitions of certainty in evidence, one for the former setting and one for the latter (Box 1) [1]. The process for setting thresholds for fully contextualized ratings of certainty may sometimes be challenging, but the need is clear. The process for setting thresholds for settings that are not fully contextualized is less clear. We will now address this issue.

### 3.2. Noncontextualized or partly contextualized ratings of certainty

Table 1 presents, in addition to the fully contextualized rating of certainty, reference to noncontextualized approaches that appear in an appendix and description of a partly contextualized approach. The noncontextualized and partly contextualized approaches are relevant primarily for systematic reviews and health technology assessments.

The two noncontextualized approaches—of potential use primarily in systematic reviews and health technology assessments—do not represent guidance for applying GRADE but are part of a complete conceptualization of the certainty of evidence. One of these noncontextualized approaches presents certainty that the true effect lies within the 95% CI. A second approach focuses on our certainty that a nonnull effect is present. We include further details of these two approaches only in Appendix 1 at www.jclinepi.com.

We will illustrate the application of the approaches using the example of the decision regarding whether to use shorter or longer duration of dual antiplatelet therapy (DAPT), that is, aspirin and clopidogrel, or a related drug, in patients with coronary artery disease who have undergone placement of drug-eluting stents in their coronary arteries [6]. Critical outcomes in this case include death, myocardial infarction, serious bleeding, and stroke. Table 2 presents an evidence summary of the impact of longer duration vs. shorter duration DAPT on these four outcomes.

#### 3.2.1. Partly contextualized ratings of certainty (typically used in systematic reviews and health technology assessments): ranges of magnitude of effect

A partly contextualized option is to rate our certainty in a specific magnitude of effect (Table 1). For instance, we could consider whether the point estimate for a single outcome, were it accurate, represents a trivial, small, moderate, or large effect. We could then rate our certainty that the true effect for this outcome, expressed in absolute terms, lies within the boundaries of whatever we consider the range of a trivial, small, medium or large effect. This is completely analogous with our prior discussion of thresholds, but now we have two: one that represents the upper, and one the lower, limit of the designations small, medium,
and large. This approach is likely to be particularly relevant for health technology assessments, or for systematic reviewers who believe the usefulness of their review will be enhanced by providing “plain language” to specify the magnitude of effect.

The challenge here is the specification of what we will consider trivial, small, moderate, and large effects. This is likely to differ across outcomes. People may, for example, consider a reduction in deaths of 6 in 100 patients (e.g., from 20% to 14%) per year—or even less—a large effect. People are less likely to consider the same magnitude of effect as large if the outcome is much less serious (for instance, recurrent migraine headache). Ideally, in the future, consensus approaches could achieve a consensus regarding thresholds for trivial, small, moderate, or large effects for a wide variety of outcomes.

To be optimally clear, the characterization of the size of effect would also require specification of the consequences against which the effect on the chosen outcome is being traded off. For example, one could specify that there are no harms, and a small effect would then be one large enough that, given the burden of administration, one would consider it worthwhile to use the intervention. For most interventions (e.g., taking one pill a day), this would be very small and would represent the minimum of the range of a small effect. One would also require a way of specifying the upper range of what one considers a small effect, a challenge that remains in applying the approach.

Such thresholds would only apply to absolute effects (a relative risk reduction of 50% could mean a reduction from 2% to 1%, likely a small effect, or 40% to 20%, likely a large effect). If one used this approach in our example, one might specify that the effect of longer duration DAPT on mortality was a small effect (2 in 1,000) and that we have high certainty that it is small (the upper boundary of the CI is 4 in 1,000, and there is no other reason to rate down certainty).

Considering myocardial infarction, one might specify that the point estimate represents an effect that is small (a reduction of 8 in 1,000) but we have only low certainty that the effect is indeed small Table 3. The rationale for this judgment would be as follows. The CI includes 2 in 1,000 fewer, an effect that many might consider trivial, and 12 in 1,000 fewer, an effect that (arguably) many might consider moderate. In addition, however, we found serious inconsistency in results (point estimates of relative effect ranged from 0.49 to 1.08, I^2 36%), and this raises the possibility that the effect might either be trivial or no effect at all, or might be moderate. One could apply similar logic to bleeding and stroke outcomes.

When the CI of effect estimate overlaps the null effect, as in this case for the outcome of stroke (i.e., relative association measure of 1 or absolute association measure of 0), two conclusions are possible: (1) the evidence is imprecise (e.g., small number of events) and we are unable to reliably answer the question of effectiveness or (2), the evidence is precise and the intervention is in fact not effective, or the effect is trivial. To make the latter inference (no or trivial effect), the CI needs to be sufficiently narrow to exclude the threshold of whatever one considers the lower boundary of a small effect. If the CI is tight and does not cross this threshold, one can infer that the effect is null or trivial (and presumably, the intervention, with respect to that effect, is not worth considering). If the CI is wider and is not contained within this threshold, the conclusion would be that the evidence is imprecise and cannot reliably exclude a small (or if very wide moderate or even large) effect.

In this context, the question is how certain are we that the effect lies in a particular range, with boundaries on either side of an RR of 1.0? For stroke in our example, if one set boundaries at 0.70 and 1.38, certainty would be high. If one set narrower boundaries, one would rate down for imprecision. In terms of absolute effects, with a CI of 2 fewer to 2 more strokes, one would likely conclude there is a precise estimate of a trivial or null absolute effect. Making a judgment that an intervention effect is no greater than trivial will, just as for the other ranges, require some degree of contextualization—for instance, the more important the outcome, the narrower the range in which one will be willing to conclude that there is no important effect.

---

**Table 2. Longer duration vs. shorter duration DAPT after drug-eluting stents: partial evidence profile**

<table>
<thead>
<tr>
<th>Outcome</th>
<th>No. of participants (no. of studies)</th>
<th>Risk of bias</th>
<th>Inconsistency</th>
<th>Indirectness</th>
<th>Publication bias</th>
<th>Relative risk (95% CI) per 1,000 treated (95% CI) per year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total mortality</td>
<td>28,088 (9)</td>
<td>No serious limitations</td>
<td>No serious inconsistency</td>
<td>No serious indirectness</td>
<td>Undetected</td>
<td>1.19 (1.04–1.36) 2 more (0 more to 4 more)</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>28,088 (9)</td>
<td>No serious limitations</td>
<td>Serious inconsistency</td>
<td>No serious indirectness</td>
<td>Undetected</td>
<td>0.73 (0.58–0.92) 8 fewer (12 to 2 fewer)</td>
</tr>
<tr>
<td>Serious bleeding</td>
<td>26,475 (8)</td>
<td>No serious limitations</td>
<td>No serious indirectness</td>
<td>No serious indirectness</td>
<td>Undetected</td>
<td>1.63 (1.34–1.99) 6 more (3 more to 10 more)</td>
</tr>
<tr>
<td>Stroke</td>
<td>28,088 (9)</td>
<td>No serious limitations</td>
<td>No serious indirectness</td>
<td>No serious indirectness</td>
<td>Undetected</td>
<td>0.99 (0.71–1.37) 0 more (2 fewer to 2 more)</td>
</tr>
</tbody>
</table>

**Abbreviations:** CI, confidence interval; DAPT, dual antiplatelet therapy.

a The fifth domain, imprecision, is not presented in this table because the assessment of imprecision is dependent on the chosen threshold or range (Table 1).
3.3. Applying fully contextualized ratings (typically used in clinical practice guidelines) to individual outcomes

When, typically in the setting of a guideline or recommendation, we make fully contextualized ratings, we are simultaneously weighing the benefits and harms of every important outcome. Such contextualized ratings may best be addressed at the stage of the evidence to decision, rather than at an earlier point in the decision-making process (i.e., the evidence profile or summary of findings stage). At whatever stage, a guideline panel decides to make the assessment, having separate certainty ratings for each outcome in the fully contextualized setting can inform patients, clinicians, and researchers as to where there are important gaps in medical knowledge.

The fully contextualized ratings of individual outcomes do not altogether resolve the issue of how best to rate the certainty in the net benefit. Indeed, alternative approaches to addressing certainty of evidence in the fully contextualized setting, and in particular an overall rating of certainty in net benefit, are issues of ongoing GRADE Working Group activity.

When simultaneously considering all outcomes, however, the lowest rating of certainty among the critical outcomes will generally provide an upper limit for the overall certainty in the balance between desirable and undesirable health outcomes (i.e., the net benefit). This is what GRADE currently refers to as the overall certainty of evidence [7] and thus—pending further conceptual development—it provides an interim approximation of certainty in the net benefit. Once an approach is developed for assessing certainty in the net benefit, fully contextualized ratings for individual outcomes may no longer be needed.

One approach to a fully contextualized rating of net benefit would be using a decision model, in which sensitivity analyses would highlight which outcomes are capable of tipping the model (altering the overall result from benefit to harm or vice versa) over a range of plausible values for those outcomes. The approach to fully contextualized ratings we are suggesting in this article is analogous to the decision model approach. The approach involves differential weighting of the importance of outcomes and allows a guideline panel to, without creating a decision model, address net benefit.

Making fully contextualized ratings of certainty—and indeed, deciding if one recommends for or against an intervention—requires first specifying values. The values should be those of the patients, and GRADE [8] and others provide guidance regarding how to obtain estimates of those values. The process includes a systematic review of the relevant literature [9], the experience of the topic experts in conducting shared decision making, consultation with patients and patient groups, and conduct of targeted surveys [10–12].

In the DAPT example, a guideline panel might note that they believe that typical patients would value a myocardial infarction and serious bleed similarly, place an appreciably greater value on stroke (say, three times the value of a bleed or myocardial infarction) and an even greater value on death (say, five times the value of a bleed or myocardial infarction).

### Table 3. Possible certainty ratings for myocardial infarction (MI) for longer duration vs. shorter duration DAPT

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Examples of set thresholds or ranges</th>
<th>Certainty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specified magnitude: small effect</td>
<td>The effect is small over a range of 4–11 fewer MIs per 1,000</td>
<td>We have low certainty that longer duration DAPT gives a small decrease in the incidence of MI compared with shorter duration DAPT (rating down for inconsistency and imprecision). Overall, considering typical values and preferences (equal weight to MI and serious bleeding, high importance to mortality, aversion to taking medication with minimal net benefit), we have low certainty that longer duration DAPT does not decrease MI sufficiently to outweigh the effects on survival, bleeding, and the burden associated with long-term use of additional medication. We have low certainty in the MI outcome because the overall balance between net benefit and net harm differs across the certainty range for MI (rating down for inconsistency and imprecision).</td>
</tr>
<tr>
<td>Threshold determined with considerations of all critical outcomes</td>
<td>Threshold based on the value we place on MI, bleeding, stroke, and death</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviation: DAPT, dual antiplatelet therapy.
range for all outcomes, the final recommendation may be in the opposite direction. Nevertheless, focusing on the point estimates provides a useful starting point. In this case, using the weights we have specified previously, the reduction in MI of 8 in 1,000 over 1 year suggests a recommendation in favor of longer duration DAPT, but the increase in bleeding of 6 and in death of 2 more than balances the benefit. Thus, the recommendation would be against the use of longer duration DAPT.

Having decided on the direction, let us consider the certainty rating, beginning with the outcome of death. Because, for mortality, there are no serious limitations in risk of bias, consistency, directness, or publication bias, the certainty (high or moderate) depends on the judgment of precision. One might start the process of rating precision by looking at the point estimates of outcomes other than the one under consideration (in this case, outcomes other than death). Considering the values and preferences mentioned previously (equal value on MI and bleeding) and ignoring mortality, if longer duration DAPT reduced myocardial infarctions by 8 in 1,000, increased bleeds by 6 in 1,000, and did not change the incidence of stroke, we would recommend in favor of longer duration DAPT (though barely, it would be a close call).

Now, what if longer duration DAPT actually had no effect on mortality (a risk difference of 0, at one boundary of the CI)? We would continue to recommend longer duration DAPT. What about the other end of the CI, an increase in 4 deaths. Were this the case, we would surely recommend against longer duration DAPT. Thus, because the decision differs at the opposite ends of the CI, we rate down our certainty for imprecision.

Note, making this judgment did not require the likely painful obligation of specifying the exact mortality threshold between recommending for or against. All we needed to know is that our decision differed at either end of the CI, and therefore, the threshold must lie somewhere within the CI boundaries (as it was when we considered only point estimates) in which case we needn’t rate down our certainty. Alternatively, the threshold may be within the CI boundaries (as it was when we considered uncertainty in non-MI outcomes) therefore requiring rating down for imprecision. Our experience suggests that these decisions will be easier than attempting to define an exact threshold.

### 3.4. The consideration of optimal information size

Previous GRADE writings have suggested using the optimal information size (OIS) as a possible primary item for rating imprecision [2] (i.e., considering whether the total number of participants in the included trials is more than the number of patients generated by a conventional sample size calculation for a single adequately powered trial). This way of rating imprecision is not compatible with the approaches described in the present article. However, whichever of our suggested approaches reviewers are using, they will sometimes confront large effect sizes with apparently satisfactory CIs despite modest sample size. Because such findings are untrustworthy—experience has shown that these large effects typically decrease or disappear as data accumulate—they require consideration of the event rate using OIS or closely related alternative approaches [4].

### 3.5. Presentation of certainty of evidence in the context of clinical practice guidelines

Both fully and less or noncontextualized ratings represent, in the guideline context, options for presenting certainty (in, for instance, evidence to decision tables). An advantage of choosing less-contextualized ratings is that such a presentation may be more useful for another group that wishes to adapt the guideline to a different context
specification of values and preferences\[13,14\]. Decision
Methodologists have suggested alternative quantitative
of multiple outcomes, including uncertainty in estimates.
approach.

Indeed, the advantage of choosing the fully contextual-
ized approach is that the simultaneous consideration of
all outcomes, and the implications for certainty of evidence,
determines the ultimate direction and strength of recom-
mendations. High or moderate certainty with a large
gradient between desirable and undesirable outcomes will
dictate a strong recommendation. Low certainty, or a small
gradient between desirable and undesirable outcomes, will
generally dictate a weak recommendation.

In making the decision regarding whether to present
less-contextualized or noncontextualized ratings of evi-
dence along with certainty ratings that drive the direction
and strength of their recommendations, guideline panels
may want to consider both what is optimal for the process
of coming to the recommendation, and what will be most
helpful for the ultimate consumers of the guideline. They
should specify, clearly and explicitly, which approach to
setting thresholds for effect they used.

3.6. Some limitations in the discussion

Bayesian thinking, and formal Bayesian statistics, would
provide an alternative approach to the questions we have
addressed in this article. This would be interesting to
pursue but is beyond the scope of the current discussion
because, currently, guideline developers seldom use the
approach.

Formal decision analysis based on expected utility the-
ory provides a structure for the simultaneous considera-
tion of multiple outcomes, including uncertainty in estimates.
Methodologists have suggested alternative quantitative
approaches to decision making that rely on explicit
specification of values and preferences \[13,14\]. Decision
analysis, as well as alternative quantitative approaches,
could therefore be a potential solution to the challenges
we raise in our discussion of the fully contextualized rating.
Clinical practice guidelines seldom, however, involve
formal decision analysis. Decision analysis has had even
less impact on individual patient decision making, and new-
er quantitative methods have not yet stood the test of time.
Ultimately, such approaches may provide an alternative
framework for the simultaneous consideration of all
outcomes.

4. Conclusions

This article has addressed the following question: when we
rate certainty of evidence, what exactly is it in which we are
rating our certainty. The answer we have provided is that we
are rating our certainty that the true effect lies on one side of
a particular threshold, or in a particular range. What follows
from this is the desirability for systematic review authors,
guideline panelists, and health technology assessors to specify
the threshold or ranges they are using. We have presented how
this might be done in the fully contextualized setting and pre-

dented alternatives for less-contextualized settings. Future
research can assess which approaches are most useful for
different settings and target groups, as well as how best authors
can communicate the threshold or ranges they are using.
Finally, although all our examples relate to intervention ef-
fects, the guidance that ratings of certainty should specify
the relevant thresholds underlying the judgments also applies
to questions of diagnosis and prognosis.

Acknowledgments

The authors would like to thank colleagues at the
Swedish Agency for Health Technology Assessment and
Assessment of Social Services (SBU) who, through partici-
pating in a series of seminars, contributed with valuable
input on the initial draft of the presented approaches. The
authors also thank all the GRADE Working Group members
who have contributed to the article during group discussions
at Grade Working Group meetings.

Supplementary data

Supplementary data related to this article can be found at
http://dx.doi.org/10.1016/j.jclinepi.2017.05.006.

References

et al. GRADE Equity Guidelines 3: health equity considerations in
rating the certainty of synthesized evidence. J Clin Epidemiol
[4] 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
guide line development in asthma and other diseases. Curr
[6] Spencer FA, Prasad M, Vandvik PO, Chetan D, Zhou Q, Guyatt G.
Longer- versus shorter-duration dual-antiplatelet therapy after drug-
eluting stent placement: a systematic review and meta-analysis.
et al. GRADE guidelines: 11. Making an overall rating of confidence in
effect estimates for a single outcome and for all outcomes. J Clin


