ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1283–1293

Journal of Clinical Epidemiology

GRADE guidelines 6. Rating the quality of evidence—imprecision

Gordon H. Guyatt[a,b,*], Andrew D. Oxman[c], Regina Kunz[d,e], Jan Brozek[a], Pablo Alonso-Coello[f], David Rind[g], PJ Devereaux[a], Victor M. Montori[h], Bo Freyschuss[i], Gunn Vist[c], Roman Jaeschke[b], John W. Williams Jr.[j], Mohammad Hassan Murad[h], David Sinclair[k], Yngve Falck-Ytter[l], Joerg Meerpohl[m,n], Craig Whittington[o], Kristian Thorlund[a], Jeff Andrews[p], Holger J. Schünemann[a,b]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[b]*Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[c]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, 0130 Oslo, Norway*
[d]*Academy of Swiss Insurance Medicine (asim), University Hospital Basel, Petergraben 4, CH-4031 Basel, Switzerland*
[e]*Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland*
[f]*Iberoamerican Cochrane Center-Servicio de Epidemiología Clínica y Salud Pública and CIBER de Epidemiología y Salud Pública (CIBERESP), Hospital de Sant Pau, Universidad Autónoma de Barcelona, Barcelona 08041, Spain*
[g]*Department of Medicine, Harvard Medical School, Boston, USA*
[h]*Knowledge and Encounter Research Unit, Mayo Clinic, Rochester, MN, USA*
[i]*Department of Medicine, Karolinska Institute M54, Karolinska University Hospital, 141 86 Stockholm, Sweden*
[j]*Durham VA Center for Health Services Research in Primary Care, Duke University Medical Center, Durham, NC 27705, USA*
[k]*Effective Health Care Research Consortium, Liverpool School of Tropical Medicine. Liverpool L3 5QA, UK*
[l]*Department of Medicine, Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*
[m]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany*
[n]*Division of Pediatric Hematology and Oncology, Department of Pediatric and Adolescent Medicine, University Medical Center Freiburg, 79106 Freiburg, Germany*
[o]*National Collaborating Centre for Mental Health, Centre for Outcomes Research and Effectiveness, Research Department of Clinical, Educational & Health Psychology, University College London, 1-19 Torrington Place, London WC1E 7HB, UK*
[p]*Vanderbilt Evidence-based Practice Centre, Vanderbilt University, Nashville, Tennessee*

Accepted 5 January 2011; Published online 11 August 2011

Abstract

GRADE suggests that examination of 95% confidence intervals (CIs) provides the optimal primary approach to decisions regarding imprecision. For practice guidelines, rating down the quality of evidence (i.e., confidence in estimates of effect) is required if clinical action would differ if the upper versus the lower boundary of the CI represented the truth. An exception to this rule occurs when an effect is large, and consideration of CIs alone suggests a robust effect, but the total sample size is not large and the number of events is small. Under these circumstances, one should consider rating down for imprecision. To inform this decision, one can calculate the number of patients required for an adequately powered individual trial (termed the ''optimal information size'' [OIS]). For continuous variables, we suggest a similar process, initially considering the upper and lower limits of the CI, and subsequently calculating an OIS.

Systematic reviews require a somewhat different approach. If the 95% CI excludes a relative risk (RR) of 1.0, and the total number of events or patients exceeds the OIS criterion, precision is adequate. If the 95% CI includes appreciable benefit or harm (we suggest an RR of under 0.75 or over 1.25 as a rough guide) rating down for imprecision may be appropriate even if OIS criteria are met.

Keywords: GRADE; Quality of evidence; Confidence in estimates; Imprecision; Optimal information size; Confidence intervals

1. Introduction

In five previous articles in our series describing the GRADE system of rating the quality of evidence and grading the strength of recommendations, we have described the process of framing the question, introduced GRADE's approach to quality-of-evidence rating, and described two reasons for rating down quality of evidence because of bias:

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) system has been developed by the GRADE Working Group. The named authors drafted and revised this article. A complete list of contributors to this series can be found on the journal's Web site at www.elsevier.com.

* Corresponding author. CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, Room 2C12, 1200 Main Street, West Hamilton, Ontario, Canada L8N 3Z5. Tel.: +905-527-4322; fax: +905-523-8781.

E-mail address: guyatt@mcmaster.ca (G.H. Guyatt).

0895-4356/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi: 10.1016/j.jclinepi.2011.01.012

Key Points

- GRADE's primary criterion for judging precision is to focus on the 95% confidence interval (CI) around the difference in effect between intervention and control for each outcome.
- In general, the CIs to consider are those around the absolute, rather than the relative effect.
- If a recommendation or clinical course of action would differ if the upper versus the lower boundary of the CI represented the truth, consider the rating down for imprecision.
- Even if CIs appear satisfactorily narrow, when effects are large and both sample size and number of events are modest, consider the rating down for imprecision.

study limitations and publication bias. In this article, we address another reason for rating down evidence quality: random error or imprecision.

We begin our discussion by highlighting the differences between systematic reviews and guidelines in the definitions of quality of evidence (i.e., confidence in estimates of effect) and thus in the criteria for judgments regarding precision. We then describe the key point of the article: how one can use CIs as the primary tool for judging precision (or the lack it), and how to examine the relation between CI boundaries and important effects for binary outcomes in the context of clinical practice guidelines.

Unfortunately, there are limitations of CIs; we will suggest a potential solution to the problem—the optimal information size. After summarizing our approach to evaluating precision in the context of guidelines, we apply the same logic to assessing precision in systematic reviews, the special case of low event rates, and how our approach applies to continuous variables.

2. Criteria for imprecision differ for guidelines and systematic reviews

GRADE defines evidence quality differently for systematic reviews and guidelines. For systematic reviews, quality refers to our confidence in the estimates of effect. For guidelines, quality refers to the extent to which our confidence in the effect estimate is adequate to support a particular decision.

3. Confidence intervals capture the extent of imprecision—mostly

To a large extent, CIs inform the impact of random error on evidence quality. Within the frequentist (in contrast to Bayesian) framework, the CI represents that range of results which, were an experiment repeated numerous times and the CI recalculated for each experiment, a particular proportion of the CIs (typically 95%), would include the true underlying value. Conceptually easier than this definition is to think of the CI as the range in which the truth plausibly lies.

When considering the quality of evidence, the issue is whether the CI around the estimate of treatment effect is sufficiently narrow. If it is not, we rate down the evidence quality by one level (for instance, from high to moderate). If the CI is very wide, we might rate down by two levels.

4. Guidelines: are results of a binary outcome sufficiently precise to support a recommendation?

The following example illustrates how guideline developers must consider the context of their particular recommendations in making judgments about precision. A hypothetical systematic review of randomized conrol trials (RCTs) of an intervention to prevent major strokes yields a pooled estimate of the absolute reduction in strokes of 1.3%, with a 95% CI of 0.6% to 2.0% (Fig. 1). Thus, we must treat 77 (100/1.3) patients for a year to prevent a single major stroke. The 95% CI around the number needed to treat (NNT)—50 to 167—tells us that while 77 is our best estimate, we may need to treat as few as 50 or as many as 167 people to prevent a single stroke.

Further, assume that the intervention is a drug with no serious adverse effects, minimal inconvenience, and modest cost. Under these circumstances, even a small effect would warrant a strong recommendation. For instance, we may strongly recommend the intervention were it to reduce strokes by as little as 0.5% (vertical middle line in Fig. 1)—an NNT of 200. The entire CI (0.6% to 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 therefore conclude that

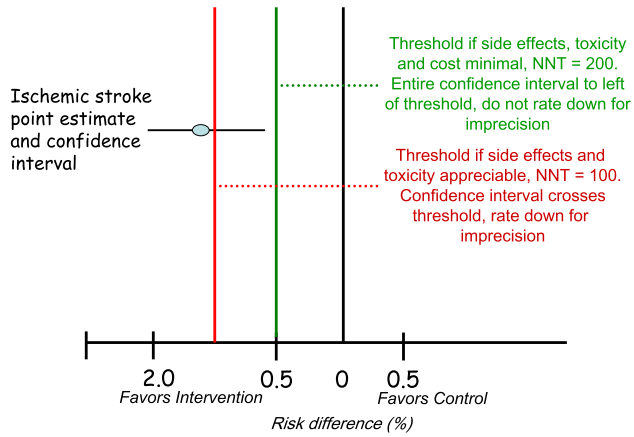

Fig. 1. Rating down for imprecision in guidelines: thresholds are key.

Box 1 The impact of undesirable consequences on precision

The hypothetical example presented in Fig. 1 and the accompanying text shows that greater levels of precision are required to support a recommendation in favor of a treatment when a large effect is required to make treatment worth the appreciable undesirable consequences. When appreciable undesirable consequences exist, CIs are more likely to span not only regions of effect that would mandate treating but also regions that would mandate not treating. Thus, the existence of appreciable undesirable consequences makes it more likely that guideline developers will rate down the evidence regarding an apparently beneficial intervention for imprecision.

Box 2 A second real world example of rating down for imprecision

Fig. 2 presents another example, a meta-analysis of trials of the use of steroids for patients in septic shock, in which a total of 511 patients died. The CI for the pooled effect (0.75 to 1.03) overlaps a relative risk (RR) of 1.0 (no effect), suggesting that a recommendation against steroids would be appropriate. Nevertheless, the boundary of the CI consistent with the largest plausible effect suggests that steroids might reduce the RR of death by as much as 25% - an effect of unequivocal importance considering typical mortality rates of 40% or more in patients with sepsis (indicating an absolute risk reduction of at least 10%). Therefore, the possibility that the RR reduction is as great as 25% would mandate rating the quality of evidence supporting a recommendation against administering steroids as moderate rather than high.

the precision of the evidence is sufficient to support a strong recommendation.

What if, however, treatment is associated with serious toxicity? Were this true, we may be reluctant to recommend treatment unless the absolute stroke reduction is at least 1% (NNT of 100—left verticle line in Fig. 1). Under these circumstances, the precision is insufficient to support a strong recommendation as the CI encompasses treatment effects smaller than this threshold and therefore does not exclude an absolute benefit appreciably less than 1%. Because the point estimate of 1.3% meets the threshold criterion, a recommendation in favor of treatment would still be appropriate, although the imprecision-generated uncertainty regarding the true effect would mandate a weak recommendation (Box 1).

5. Real world examples of the clinical decision threshold approach to precision

An RCT (the sole trial addressing the question) compared clopidogrel or aspirin in patients who have experienced a transient ischemic attack, cardiac, or peripheral

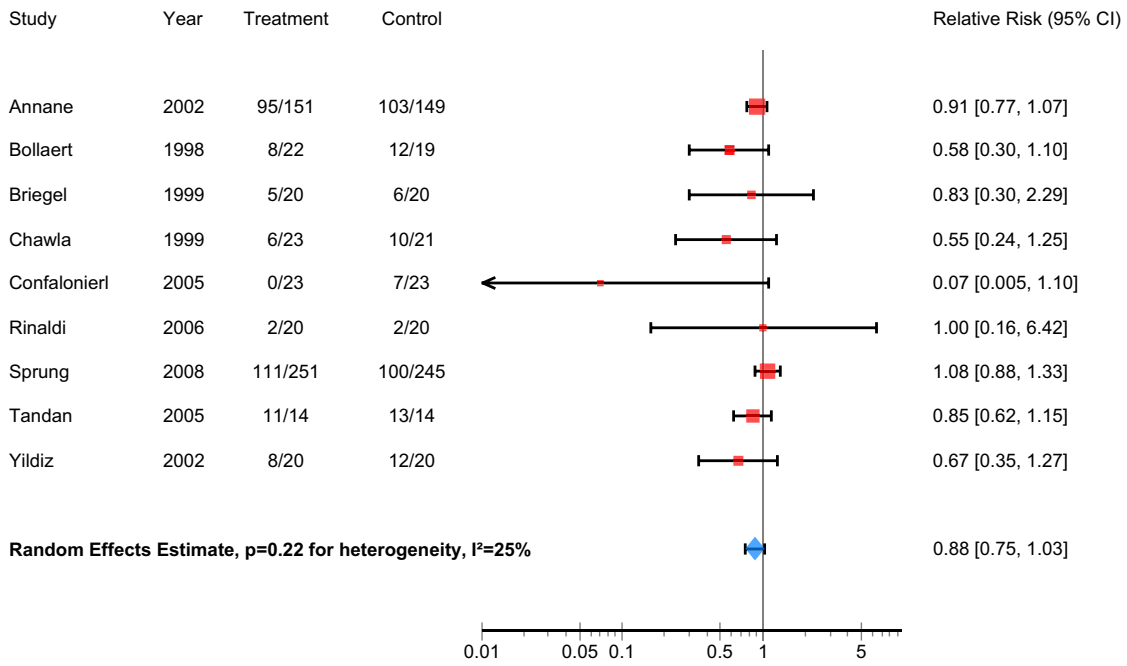

Fig. 2. Corticosteroids to reduce hospital mortality in septic shock.

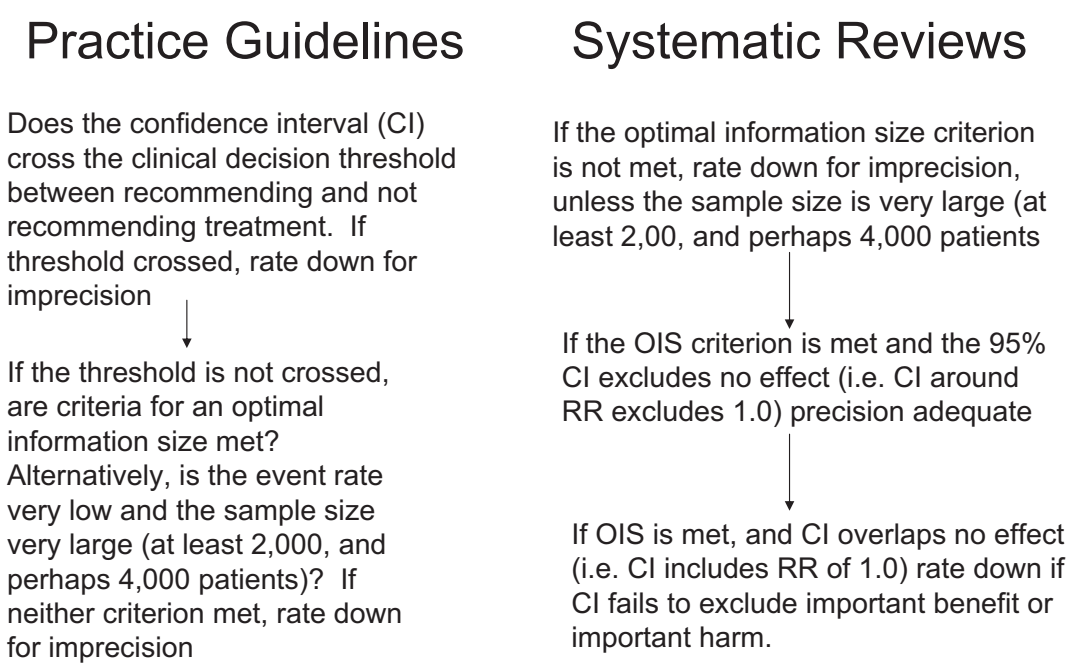

Fig. 3. Deciding whether to rate down for imprecision in guidelines and systematic reviews of binary variables.

ischemia [1]. This concealed blinded RCT enrolled 19,185 patients at risk of vascular events. Of the patients receiving clopidogrel, 939 (5.32%) experienced a major vascular event, as did 1,021 (5.83%) of those receiving aspirin. The result represents an RR of 0.91 (95% CI: 0.83, 0.99). If the CI boundary closest to no effect (a 1% relative risk reduction [RRR]) represented the true effect, guideline panels would recommend against this medication (as long, at least, as clopidogrel remains costly). Thus, despite the huge sample size and very large number of events, trial results are insufficiently precise to support a treatment recommendation, and rating down quality by one level for imprecision is mandated. Box 2 presents another example of rating down for imprecision.

The reasoning in the examples above relies on value-and-preference judgments. A number of factors will influence the decision, including the importance of the outcome (e.g., mortality vs. improving symptoms), the adverse effects, the burden to patient, and perhaps resource use and the difficulty of introducing the intervention into practice. Ideally, these judgments would reflect average judgments of an informed public. Unfortunately, empirical evidence of average public values and preferences is limited. This argues for guideline panels being completely explicit—and as quantitative as possible—about their value–and-preference judgments.

In summary, when guideline developers consider imprecision, the first step is to determine whether CIs cross a clinical decision threshold that dictates recommending versus not recommending an intervention (Fig. 3). The remainder of this article addresses the limitations of CIs, a potential solution to these limitations, and the limitations of the solution. Readers can consider these issues secondary to the primary criteria that we have thus far addressed.

6. Confidence intervals can be misleading because of fragility

The clinical decision threshold criterion is not completely sufficient to deal with issues of precision. The reason is that Cis may appear robust, but small numbers of events may render the results fragile (see Box 3 for an example).

7. The danger of initial trials with impressive positive results

Simulation studies [3] and empirical evidence [4,5] suggest that trials stopped early for benefit overestimate treatment effects. Investigators have tested thousands of questions in RCTs, and perhaps hundreds of questions are being addressed in ongoing trials. Some early trials addressing a particular question will, particularly if small, substantially overestimate the treatment effect. A systematic review of these early trials will also generate a spuriously large effect estimate. If a false large effect estimate from a systematic review stifles subsequent investigation, the situation is analogous to a single RCT stopped early for apparent benefit.

Another way of thinking of the limitations of CIs is in terms of prognostic balance. CIs assume all patients are at

Box 3 An example of fragility

Consider a randomized trial of β blockers in 112 patients undergoing surgery for peripheral vascular diseases that fulfilled preplanned O'Brien−Fleming criteria for early stopping [2]. Of 59 patients given bisoprolol, 2 suffered a death or nonfatal myocardial infarction, as did 18 of 53 control patients. Despite a total of only 20 events, the 95% CI around the RR (0.02 to 0.41) excludes all but a large treatment effect. The CI suggests that the smallest plausible effect is a 59% RRR. Were this the case, we would certainly administer treatment. Thus, according to criteria discussed up to now, a recommendation based on this result would be deemed to have adequate precision.

There are reasons to doubt the estimate of the magnitude of effect from this trial. First, it is much larger than what we might expect on the basis of β blockers effects in a wide variety of other situations. Second, the study was terminated early on the basis of the large effect. Third, concluding that an RRR less than 59% is implausible on the basis of only 20 events violates common sense: intuitively, we have a sense of the fragility of these results. Our intuitive skepticism is justified: if one moves just five events from the control to the intervention group, the results lose their statistical significance, and the new point estimate (an RRR of 52%) is outside of the original CI.

the same risk—an assumption that is false. Randomization will ameliorate the problem of varying prognosis by balancing prognosis in intervention and control groups. We can be confident that we have achieved this prognostic balance, however, only if sample sizes are large. Impressive treatment effects in the presence of small sample size may well—even in RCTs—be because of prognostic imbalance.

These considerations argue for skepticism regarding evidence summaries that generate apparent benefits, or harms, of therapy with what appear to be satisfactorily narrow CIs on the basis of small trials with relatively few events. Examples of meta-analyses generating apparent beneficial or harmful effects refuted by subsequent larger trials, include magnesium for mortality reduction after myocardial infarction [6,7], angiotensin-converting-enzyme inhibitors for reducing the incidence of diabetes [8,9], β blockade for cardiovascular mortality reduction in patients undergoing noncardiac surgery [10,11], nitrates for mortality reduction in myocardial infarction [12,13], aspirin for reduction of pregnancy-induced hypertension [14,15], albumin for mortality reduction in the critically ill [16,17], and a number of mental health interventions [18].

Box 4 Applying the optimal information size using total sample size or number of events

A systematic review of flavonoids for treatment of hemorrhoids examined the outcome of failure to achieve an important symptom reduction [20]. In calculating the OIS, the authors chose a conservative α of 0.01 and RRR (20%), a β of 0.2, and a control event rate of 50%. The authors found that the OIS was marginally larger than the total sample size included (1,194 vs. 1,102 patients).

A more dramatic example comes from a systematic review and meta-analysis of fluoroquinolone prophylaxis for patients with neutropenia [21]. Only one of eight studies that contributed to the meta-analysis met conventional criteria for statistical significance, but the pooled estimate suggested an impressive and robust reduction in infection-related mortality with prophylaxis (RR: 0.38; 95% CI: 0.21, 0.69). The total number of events, however, was only 69 and the total number of patients 1,022. Considering the control event rate of 6.9% and setting α of 0.05, β of 0.02, and RRR of 25% results in an OIS of 6,400 patients. This meta-analysis, therefore, fails to meet OIS criteria, and rating down for imprecision may be warranted.

8. Addressing the vulnerability of CIs: the optimal information size

The reasoning above suggests the need for, in addition to CIs, another criterion for adequate precision. We suggest the following: if the total number of patients included in a systematic review is less than the number of patients generated by a conventional sample size calculation for a single adequately powered trial, consider the rating down for imprecision. Authors have referred to this threshold as the "optimal information size" (OIS) [19]. Many online calculators for sample size calculation are available—you can find one simple one at http://www.stat.ubc.ca/~rollin/stats/ssize/b2.html.

Box 4 presents examples of application of the OIS.

As an alternative to calculating the OIS, review and guideline authors can also consult a figure to determine the OIS. Fig. 4 presents the required sample size (assuming α of 0.05, and β of 0.2) for RRR of 20%, 25%, and 30% across varying control event rates. For example, if the best estimate of control event rate was 0.2 and one specifies an RRR of 25%, the OIS is approximately 2,000 patients.

Power is, however, more closely related to number of events than to sample size. Fig. 5 presents the same relationships using total number of events across all studies in both treatment and control groups instead of total

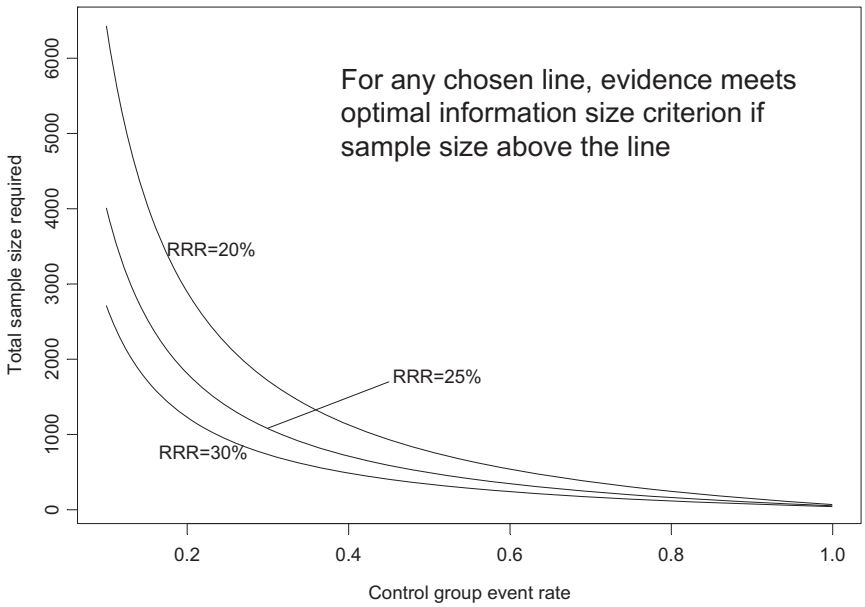

Fig. 4. Optimal information size given α of 0.05 and β of 0.2 for varying control event rates and relative risks.

number of patients. Using the same choices as in the prior paragraph (control event rate 0.2 and RRR 25%), one requires approximately 325 events to meet OIS criteria.

We have suggested using RRRs of 20% to 30% for calculating OIS. The choice of RRR is a matter of judgment, and there may be instances in which compelling prior information would suggest choosing a larger value for the RRR for the OIS calculation.

If guideline panels are disinclined to calculate their own OIS (although calculating is preferable), they can use Figs. 4 and 5 to determine OIS. In doing so, they will note the sample size implications in Table 1.

9. Low event rates with large sample size: an exception to the need for OIS

In the criteria we have so far offered, our focus has been on relative effects. When event rates are very low, CIs around relative effects may be wide, but if sample sizes are sufficiently large, it is likely that prognostic balance has indeed been achieved, and rating down for imprecision becomes inappropriate.

For example, consider a systematic review of artemether–lumefantrine versus Amodiaquine plus sulfadoxine–pyrimethamine for treating uncomplicated malaria. For

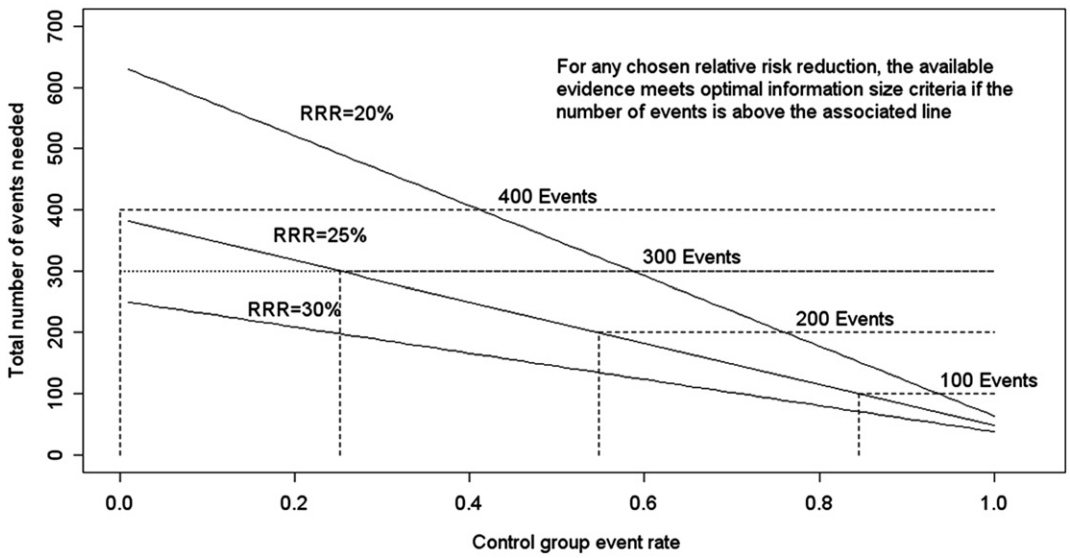

Fig. 5. Optimal information size presented as number of events given α of 0.05 and β of 0.20 for varying control event rates and RRR of 20%, 25% and 30%. *Abbreviation*: RRR, relative risk reduction.

Table 1. Optimal information size implications from Fig. 5

Total number of events	RRR (%)	Implications for meeting OIS threshold
100 or less	≤30	Will almost never meet threshold whatever control event rate
200	30	Will meet threshold for control group risks of ~ 25% or greater
200	25	Will meet threshold for control group risks of ~ 50% or greater
200	20	Will meet threshold only for control group risks of ~ 80% or greater
300	≥30	Will meet threshold
300	25	Will meet threshold for control group risks of ~ 25% or greater
300	20	Will meet threshold for control group risks of ~ 60% or greater
400 or more	≥25	Will meet threshold for any control group risks
400 or more	20	Will meet threshold for control group risks of ~ 40% or greater

Abbreviations: RRR, relative risk reduction; OIS, optimal information size.

serious adverse events (SAEs), the authors calculated a RR of 1.08 (95% CI: 0.56, 2.08). They judged this CI sufficiently wide to rate down quality two levels (from high to low) for imprecision.

There were, however, only 34 SAEs in over 2,700 patients, corresponding to event rates of 1.2 and 1.3% in the two groups. Of these events, 2 were deaths, 2 severe anemias, and the remainder febrile seizures and elevated liver function. In absolute terms, the difference between groups is 1 event per thousand patients with a CI from 6 in 1,000 fewer to 14 in 1,000 more. Particularly considering that very few of these SAEs were associated with long-term morbidity, focusing on the CI around absolute versus relative effects would lead one to reject rating down quality two levels for imprecision, and possibly not rate down for imprecision at all. Box 5 presents a second example of this issue.

Box 5 An example of low event rates and appropriate focus on absolute rather than relative effects

A systematic review of seven randomized trials of angioplasty versus carotid endarterectomy for cerebrovascular disease found that a total of 16 of 1,482 (1.1%) patients receiving angioplasty died, as did 19 of 1,465 (1.3%) undergoing endarterectomy [22]. Looking at the 95% CI (0.43–1.66) around the point estimate of the RR (0.85), the results are apparently consistent with substantial benefit and substantial harm, suggesting the need to rate down for imprecision.

The absolute difference, however, suggest a different conclusion. As it turns out, the absolute difference in death rates between the two procedures is almost certainly very small (absolute difference of 0.2% with a 95% CI ranging from −0.5% to 1.0%). Setting a clinical decision threshold boundary of 1% absolute difference (the smallest difference important to patients), the results of the systematic review exclude an important difference favoring either procedure. If one accepted this clinical decision threshold as appropriate, one would not rate down for imprecision. One could argue that a difference of less than 1% could be important to patients: if so, one would rate down for imprecision, even after considering the CI around the absolute difference.

The decision regarding the magnitude of effect that would be important is a matter of judgment. When control rates are sufficiently low, CIs around relative effects can appear very wide, but CIs around absolute effects will nevertheless be narrow. Thus, although one would intuitively rate down for imprecision considering only the CI around the relative effect, consideration of the CI around the absolute effect may lead to an appropriate conclusion that precision is adequate. Note: The inference of unimportance requires a low incidence of events over the desirable duration of follow-up; short follow-up will generate a low incidence of events that may be misleading.

Similarly, if sample sizes are sufficiently large, one need not apply the OIS criteria when results show an apparent treatment effect with a satisfactory CI. Box 6 provides an example.

10. Rating precision for binary variables in guidelines: summary and conclusions

Fig. 3 summarizes our approach to rating down quality of evidence for imprecision in guidelines. Initially, guideline developers consider whether the boundaries of the CI are on the same side of their decision-making threshold. If the answer is no (i.e., the CI crosses the threshold), one rates down for imprecision irrespective of the where the point estimate and CIs lie.

If the answer is yes (both boundaries of the CI lie on one side of the clinical decision threshold), one determines whether the OIS criterion is met. If it is met, imprecision is not a concern. If it is not met, guideline authors should consider rating down for imprecision. If event rates are very low, however, CIs around absolute effects are narrow and, if sample size is large, rating down for imprecision is unnecessary.

11. Standards for adequate precision of binary variables in systematic reviews: application of the OIS

Authors of systematic reviews should not rate down quality on the basis of the trade-off between desirable and

Box 6 No need to rate down for imprecision when sample sizes are very large

A meta-analysis of randomized trials of β blockade for preventing cardiovascular events in patients undergoing noncardiac surgery [23] suggested a doubling of the risk of strokes with β blockers (RR: 2.22; 95% CI: 1.39, 3.56; Fig. 6). Most trials in this meta-analysis do not suffer from important limitations, the evidence is direct and consistent, and publication bias is undetected. One would consider the lower boundary of the CI (an increase in RR of 39%) adequate precision if one believed that most patients would be reluctant to use β blockers with an increase in RR of stroke of 39%. These considerations suggest that we have high-quality evidence that β blockers increase the risk of stroke.

The total number of events (75), however, appears insufficient, an inference that is confirmed with an OIS calculation (α 0.05, β 0.20, using the β-blocker group's 1% event rate as the control, and Δ 0.25, total sample size 43,586 in comparison to the 10,889 patients actually enrolled). The guidelines we have suggested would, therefore, mandate rating down quality for imprecision.

With a sample size of over 5,000 patients per group, however, it is very likely that randomization has succeeded in creating prognostic balance. If that is true, β blockers really do increase the risk of stroke. Not rating down for imprecision in this situation is therefore appropriate. Preliminary information suggests that for low baseline risk contexts (<5%) one will be safe with regard to prognostic balance with a total of 4,000 patients (2,000 patients per group). Availability of this number of patients would mandate not rating down for imprecision despite not meeting the OIS criterion.

undesirable consequences: it is not their job to make value and preference judgments. Therefore, in judging precision, they should not focus on the threshold that represents the basis for a management decision. Rather, they should consider the OIS. If the OIS criterion is not met, they should rate down for imprecisions unless the sample size is very large. If the criterion is met, and the 95% CI around an effect excludes 1.0 (i.e., the results show a statistically significant difference), there is no need to rate down for imprecision (Fig. 3). To be of optimal use to guideline developers, a systematic review may point out what thresholds of benefit would still mandate rating down for imprecision.

12. Systematic reviews of binary variables: meeting threshold OIS may not ensure precision

Although satisfying the OIS threshold in the presence of a CI excluding no effect indicates adequate precision, the same is not true when the point estimate fails to exclude no effect. Consider, for instance, the systematic review of β blockers in noncardiac surgery mentioned previously [23]. For total mortality, with 295 deaths and a total sample size of over 10,000, the point estimate and 95% CI for the RR with β blockers were 1.24 (95% CI: 0.99, 1.56). Despite the large sample size and number of events, one might be reluctant to conclude precision is adequate when a small reduction in mortality with β blockers, as well as an increase of 56%, remain plausible.

This example suggests that when the OIS criteria are met, and the CI includes the null effect, systematic review authors should consider whether CIs include appreciable benefit or harm. Reviewers should use their judgment in deciding what constitutes appreciable benefit and harm and provide a rationale for their choice. If reviewers fail to find a compelling rationale for a threshold, our suggested default threshold for appreciable benefit and harm that warrants rating down is an RRR or RR increase of 25% or more.

For another example, consider the systematic review of steroids for reducing hospital mortality in sepsis that we described earlier (Fig. 2). The total number of events is 511; this easily meets OIS, even using a 20% RRR threshold (given a control event rate of 40% or more) (Fig. 5). The CI around the RR crosses 1.0, and the upper boundary of the CI represents a 25% RRR. Given that this 25% RRR represents a 10% absolute risk reduction, systematic review authors might well conclude that rating down for imprecision is appropriate.

13. Rating down two levels for imprecision

When there are very few events and CIs around both relative and absolute estimates of effect that include both appreciable benefit and appreciable harm, systematic reviewers and guideline developers should consider rating down the quality of evidence by two levels. For example, a systematic review of the use of probiotics for induction of remission in Crohn's disease found a single randomized trial that included 11 patients [24]. Of the treated patients, four of five achieved remission; this was true of five of six of the control patients. The point estimate of the risk ratio (0.96) suggests no difference, but the CI includes a reduction in likelihood of remission of almost half, or an increase in the likelihood of over 50% (95% CI: 0.56, 1.69).

14. Standards for adequate precision in systematic reviews of continuous variables

Review and guideline authors can calculate the OIS for continuous variables in exactly the same way they can for binary variables by specifying the α and β errors (we have suggested 0.05 and 0.2) and the Δ, and choosing an appropriate standard deviation from one of the relevant studies. For instance, a systematic review suggests that

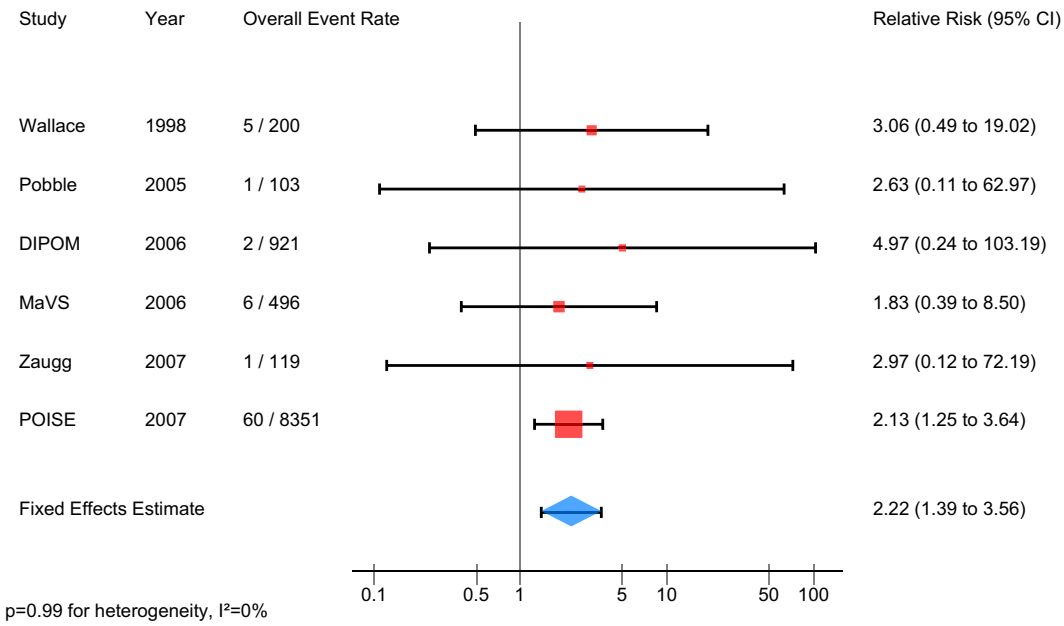

Fig. 6. Meta-analysis of beta blockers in noncardiac surgery: outcome and stroke.

corticosteroid administration decreases the length of hospital stay in patients with exacerbations of chronic obstructive pulmonary disease (COPD) by 1.42 days (95% CI: 0.65, 2.2) [25].

Choosing a Δ of 1.0 (implying a judgment that reductions in stay of more than a day are important) and using the standard deviation associated with hospital stay in the four relevant studies (3.4, 4.5, and 4.9) yield corresponding required total sample sizes of 364, 636, and 754. The 602 patients available for this analysis do not therefore meet more rigorous criteria for OIS, and one would consider rating down for imprecision.

Note that whether one will rate down for imprecision is dependent on the choice of the difference one wishes to detect. Had we chosen a smaller difference (say 0.5 days) that we wished to detect, the sample size of the studies would have been unequivocally insufficient. Had we chosen a larger value (say 1.5 days) the sample size of 602 would have comfortably met OIS criteria. As usual, the merit of the GRADE approach is not that it ensures agreement between reasonable individuals but ensures the explicitness of the judgments being made.

A particular challenge in calculating the OIS for continuous variables arises when studies have used different instruments to measure a construct, and the pooled estimate is calculated using a standardized mean difference. Systematic review and guideline authors will most often face this situation when dealing with patient-reported outcomes, such as quality of life. In this context, we suggest authors choose one of the available instruments (ideally, one in which an estimate of the minimally important difference is available) and calculate an OIS using that instrument.

Because it may give false reassurance, we hesitate to offer a rule-of-thumb threshold for the absolute number of patients required for adequate precision for continuous variables. For example, using the usual standards of α (0.05) and β (0.20), and an effect size of 0.2 standard deviations, representing a small effect, requires a total sample size of approximately 400 (200 per group)—a sample size that may not be sufficient to ensure prognostic balance.

Nonetheless, whenever there are sample sizes that are less than 400, review authors and guideline developers should certainly consider rating down for imprecision. In future, statistical simulations may provide the basis for a robust rule of thumb for continuous outcomes. The limitations of an arbitrary threshold sample size suggest the advisability of addressing precision by calculation of the relevant OIS for each continuous variable.

As is true for binary outcomes, one might consider rating down for imprecision, even if the OIS threshold is met, when the CI overlaps no effect but includes important benefit or important harm. Here again, authors must make the judgment regarding what is important. This is essentially the same judgment required for the OIS calculation—the difference one seeks to detect, 1.0 days in the example above.

15. Standards for adequate precision in guidelines addressing continuous variables

Considerations of rating down quality because of imprecision for continuous variables follow the same logic as for binary variables. The process begins by rating down the quality for imprecision if a recommendation would be altered if the lower versus the upper boundary of the CI

Box 7 Dealing with close call decisions

Our discussion has highlighted that guideline developers and systematic review authors will, not infrequently, face borderline decisions. While we have chosen binary categorical decisions (e.g., rate down for imprecision or do not rate down), the underlying quality-of-evidence concepts (in this case, imprecision) are actually continua. In situations in which differing criteria would lead to different decisions regarding rating down, it is very likely that the extent of the problem (in this case, the imprecision) is near the threshold. When it comes time to make the final judgment of quality of evidence considering other quality criteria (e.g., study limitations, consistency, directness), review and guideline authors should note if a particular decision (in this case, the decision about rating down for imprecision) was a close call. When considering all the issues that bear on quality of evidence, rating down would be more likely if the degree of imprecision was unequivocally problematic than if it were near the threshold between rating down for quality and not rating down.

For instance, assume that in the steroids for reducing length-of-stay example, we not only had a close call for rating down for imprecision but also had a close call for risk of bias. If the evidence met all other quality criteria, we would certainly rate down one level to moderate (two borderline serious limitations) but not two levels to low (because the decision to rate down was borderline in both cases and thus of limited impact on quality).

represented the true underlying effect. If the data withstand this test, but the evidence fails to meet the OIS standard, guideline authors should consider rating down the quality of evidence.

For instance, in the review of corticosteroids for exacerbations of COPD to which we referred previously, the lower boundary of the CI around the reduction in days in hospital was 0.65 days. If the effect was really this small, would one still recommend the administration of corticosteroids?

In the context of a guideline (as opposed to a systematic review), the decision requires consideration of the full context, including other outcomes. As it turns out, steroids also reduce the likelihood of ''treatment failure'' (variably defined) during inpatient or outpatient follow-up (RR: 0.54; 95% CI: 0.41, 0.71). The best estimate of likelihood of symptomatic deterioration in those not treated with steroids is approximately 30%. By administering steroids to these patients, we can reduce this 30% risk to 16% (30−[0.54 × 30]), a difference of 14%, and the effect is unlikely to be less than 9% (30−[0.71 × 30]).

Adverse effects were poorly reported in the studies. The only consistently reported problem was hyperglycemia, which was increased almost sixfold, representing an absolute increase of 15% to 20%. The extent to which this hyperglycemia had consequences important to patients is uncertain.

One possible conclusion from this information is that, given the magnitude of reduction in deterioration and lack of evidence suggesting important adverse effects, a benefit of even 0.65 days of reduced average hospitalization would warrant steroid administration. If this were their conclusion, a guideline panel would proceed to consider whether the evidence meets the OIS criterion as presented in the previous section.

16. Conclusion

Consideration of the impact of imprecision on quality of evidence is a complex matter (Box 7). Subsequent empirical studies may lead GRADE to modify the criteria we have suggested here. Understanding the issues will allow systematic review authors and guideline developers to judiciously apply the guidance we have suggested.

References

[1] A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events (CAPRIE). CAPRIE Steering Committee. Lancet 1996;348:1329−39.
[2] Poldermans D, Boersma E, Bax JJ, et al. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. N Engl J Med 1999;341:1789−94.
[3] Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. Control Clin Trials 1989;10:209S−21S.
[4] Montori VM, Devereaux PJ, Adhikari NK, et al. Randomized trials stopped early for benefit: a systematic review. JAMA 2005;294:2203−9.
[5] Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 2010;303:1180−7.
[6] Teo KK, Yusuf S, Collins R, et al. Effects of intravenous magnesium in suspected acute myocardial infarction: overview of randomised trials. BMJ 1991;303:1499−503.
[7] ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction. ISIS-4 (Fourth International Study of Infarct Survival) Collaborative Group. Lancet 1995;345:669−85.
[8] Abuissa H, Jones PG, Marso SP, et al. Angiotensin-converting enzyme inhibitors or angiotensin receptor blockers for prevention of type 2 diabetes: a meta-analysis of randomized clinical trials. J Am Coll Cardiol 2005;46:821−6.
[9] Bosch J, Yusuf S, Gerstein HC, et al. Effect of ramipril on the incidence of diabetes. N Engl J Med 2006;355:1551−62.
[10] Devereaux PJ, Beattie WS, Choi PT, et al. How strong is the evidence for the use of perioperative beta blockers in non-cardiac surgery? Systematic review and meta-analysis of randomised controlled trials. BMJ 2005;331:313−21.
[11] Bangalore S, Wetterslev J, Pranesh S, et al. Perioperative beta blockers in patients having non-cardiac surgery: a meta-analysis. Lancet 2008;372:1962−76.

[12] Yusuf S, Collins R, MacMahon S, et al. Effect of intravenous nitrates on mortality in acute myocardial infarction: an overview of the randomised trials. Lancet 1988;1:1088−92.

[13] GISSI-3: effects of lisinopril and transdermal glyceryl trinitrate singly and together on 6-week mortality and ventricular function after acute myocardial infarction. Gruppo Italiano per lo Studio della Sopravvivenza nell'infarto Miocardico. Lancet 1994;343:1115−22.

[14] Imperiale TF, Petrulis AS. A meta-analysis of low-dose aspirin for the prevention of pregnancy-induced hypertensive disease. JAMA 1991;266:260−4.

[15] CLASP: a randomised trial of low-dose aspirin for the prevention and treatment of pre-eclampsia among 9364 pregnant women. CLASP (Collaborative Low-dose Aspirin Study in Pregnancy) Collaborative Group. Lancet 1994;343:619−29.

[16] Human albumin administration in critically ill patients: systematic review of randomised controlled trials. Cochrane Injuries Group Albumin Reviewers. BMJ 1998;317:235−40.

[17] Finfer S, Bellomo R, Boyce N, et al. A comparison of albumin and saline for fluid resuscitation in the intensive care unit. N Engl J Med 2004;350:2247−56.

[18] Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. J Clin Epidemiol 2004;57:1124−30.

[19] Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials 1997;18:580−93. discussion 661−666.

[20] Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, et al. Meta-analysis of flavonoids for the treatment of haemorrhoids. Br J Surg 2006; 93:909−20.

[21] Gafter-Gvili A, Fraser A, Paul M, et al. Meta-analysis: antibiotic prophylaxis reduces mortality in neutropenic patients. Ann Intern Med 2005;142:979−95.

[22] Murad MH, Flynn DN, Elamin MB, et al. Endarterectomy vs stenting for carotid artery stenosis: a systematic review and meta-analysis. J Vasc Surg 2008;48:487−93.

[23] Devereaux PJ, Yang H, Yusuf S, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet 2008;371: 1839−47.

[24] Butterworth AD, Thomas AG, Akobeng AK. Probiotics for induction of remission in Crohn's disease. Cochrane Database of Systematic Reviews 2008; Issue 3. Art. No.: CD006634. doi:10.1002/14651858.CD006634.pub2.

[25] Quon BS, Gan WQ, Sin DD. Contemporary management of acute exacerbations of COPD: a systematic review and metaanalysis. Chest 2008;133:756−66.