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

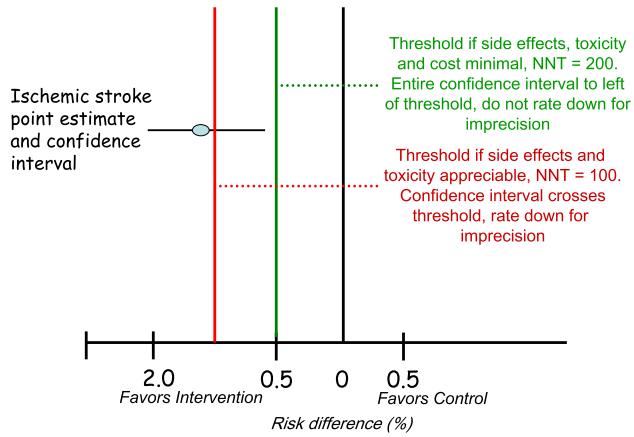

Fig. 1. Rating down for imprecision in guidelines: thresholds are key.