ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 407–415

Journal of Clinical Epidemiology

GRADE guidelines: 4. Rating the quality of evidence—study limitations (risk of bias)

Gordon H. Guyatt[a,*], Andrew D. Oxman[b], Gunn Vist[b], Regina Kunz[c], Jan Brozek[a], Pablo Alonso-Coello[d], Victor Montori[e], Elie A. Akl[f], Ben Djulbegovic[g,h,i], Yngve Falck-Ytter[j], Susan L. Norris[k], John W. Williams Jr.[l], David Atkins[m], Joerg Meerpohl[n,o], Holger J. Schünemann[a]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[b]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St. Olavs plass, 0130 Oslo, Norway*
[c]*Academy of Swiss Insurance Medicine, University Hospital Basel; and Basel Institute of Clinical Epidemiology and Biostatistics, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland*
[d]*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*
[e]*Knowledge and Encounter Research Unit, Mayo Clinic, Rochester, MN, USA*
[f]*Department of Medicine, State University of New York at Buffalo, NY, USA*
[g]*Center for Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL 33612, USA*
[h]*Department of Hematology, H. Lee Moffitt Cancer Center & Research Institute, 12901 Bruce B. Downs Boulevard, MDC02, Tampa, FL 33612, USA*
[i]*Department of Health Outcomes and Behavior, H. Lee Moffitt Cancer Center & Research Institute, 12901 Bruce B. Downs Boulevard, MDC02, Tampa, FL 33612, USA*
[j]*Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*
[k]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, USA*
[l]*Duke University Medical Center and Durham Veterans Affairs Center for Health Services Research in Primary Care Durham, NC 27705, USA*
[m]*QUERI Program, Office of Research and Development, Department of Veterans Affairs, Washington, DC, USA*
[n]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany*
[o]*Division of Pediatric Hematology and Oncology, Department of Pediatric and Adolescent Medicine, University Medical Center Freiburg, 79106 Freiburg, Germany*

Accepted 30 July 2010

Abstract

In the GRADE approach, randomized trials start as high-quality evidence and observational studies as low-quality evidence, but both can be rated down if most of the relevant evidence comes from studies that suffer from a high risk of bias. Well-established limitations of randomized trials include failure to conceal allocation, failure to blind, loss to follow-up, and failure to appropriately consider the intention-to-treat principle. More recently recognized limitations include stopping early for apparent benefit and selective reporting of outcomes according to the results. Key limitations of observational studies include use of inappropriate controls and failure to adequately adjust for prognostic imbalance. Risk of bias may vary across outcomes (e.g., loss to follow-up may be far less for all-cause mortality than for quality of life), a consideration that many systematic reviews ignore. In deciding whether to rate down for risk of bias—whether for randomized trials or observational studies—authors should not take an approach that averages across studies. Rather, for any individual outcome, when there are some studies with a high risk, and some with a low risk of bias, they should consider including only the studies with a lower risk of bias.

Keywords: GRADE; quality of evidence; risk of bias; confidence in estimates; blinding; concealment

1. Introduction

In three 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 and introduced GRADE's approach to rating the quality of evidence. This fourth article deals with one of the five categories of reasons for rating down the quality of evidence, study limitations (risk of bias).

The 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 of Clinical Epidemiology* Web site.

* 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
doi: 10.1016/j.jclinepi.2010.07.017

Key points

- In the GRADE approach, both randomized trials (which start as high quality evidence) and observational studies (which start as low quality evidence) can be rated down if relevant evidence comes from studies that suffer from a high risk of bias.
- Risk of bias can differ across outcomes when, for instance, each outcome is informed by a different subset of studies (e.g. mortality from some trials, quality of life from others).
- Current systematic reviews are often limited in their usefulness for guidelines because they rate risk of bias by studies across outcomes rather than by outcome across studies.

2. Rating down quality for risk of bias

Both randomized controlled trials (RCTs) and observational studies may incur additional risk of misleading results if they are flawed in their design or conduct—what other publications refer to as problems with ''validity'' or ''internal validity'' and we label ''study limitations'' or ''risk of bias.''

3. Study limitations in randomized trials

Readers can refer to many authoritative discussions of the study limitations that often afflict RCTs (Table 1). Two of these discussions are particularly consistent with GRADE's conceptualization, which include a focus on outcome specificity (i.e., the focus of risk of bias is not the individual study but rather the individual outcome, and quality can differ across outcomes in individual trials, or a series of trials [1,2]). We shall highlight three of the criteria in Table 1. The importance of the first of these, stopping early for benefit, has only recently been recognized. Recent evidence has also emerged regarding the second, selective outcome reporting [3,4]. Furthermore, the positioning of selective outcome reporting in taxonomies of bias can be confusing. Some may intuitively think it should be categorized with publication bias, rather than as an issue of risk of bias within individual studies. Finally, we highlight loss to follow-up because it is often misunderstood.

Before we do so, however, we note one additional issue. Recent evidence suggests that bias associated with lack of blinding and lack of concealment may be greater in trials with subjective outcomes [5]. Systematic review authors and guideline developers should consider this evidence when making decisions about rating down quality for risk of bias.

4. Stopping early for benefit

Theoretical consideration [6], simulations [7], and empirical evidence [8] all suggest that trials stopped early for benefit overestimate treatment effects. The most recent empirical work suggests that in the real world, formal stopping rules do not reduce this bias, that it is evident in stopped early trials with less than 500 events and that on average the ratio of relative risks in trials stopped early vs. the best estimate of the truth (trials not stopped early) is 0.71 [9].

Because in most cases the major contributor to the overestimation of treatment effects in trials stopped early for benefit is chance, including stopping early as a source of bias is questionable. Nevertheless, the presence of stopped early trials, particularly when they contribute substantial weight in a meta-analysis, should alert systematic review authors and guideline developers to the possibility of a substantial overestimate of treatment effect. Systematic reviews should provide sensitivity analyses of results including and excluding studies that stopped early for benefit; if estimates differ appreciably, those restricted to the trials that did not stop early should be considered the more credible. When evidence comes

Table 1
Study limitations in randomized trials

1. Lack of allocation concealment
   Those enrolling patients are aware of the group (or period in a crossover trial) to which the next enrolled patient will be allocated (major problem in ''pseudo'' or ''quasi'' randomized trials with allocation by day of week, birth date, chart number, etc)
2. Lack of blinding
   Patient, care givers, those recording outcomes, those adjudicating outcomes, or data analysts are aware of the arm to which patients are allocated (or the medication currently being received in a crossover trial)
3. Incomplete accounting of patients and outcome events
   Loss to follow-up and failure to adhere to the intention-to-treat principle in superiority trials; or in noninferiority trials, loss to follow-up, and failure to conduct both analyses considering only those who adhered to treatment, and all patients for whom outcome data are available
4. Selective outcome reporting bias
   Incomplete or absent reporting of some outcomes and not others on the basis of the results
5. Other limitations
   Stopping early for benefit
   Use of unvalidated outcome measures (e.g., patient-reported outcomes)
   Carryover effects in crossover trial
   Recruitment bias in cluster-randomized trials