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

primarily or exclusively from trials stopped early for benefit, authors should infer that substantial overestimates are likely in trials with fewer than 500 events and that large overestimates are likely in trials with fewer than 200 events [9].

5. Selective outcome reporting

When authors or study sponsors selectively report positive outcomes and analyses within a trial, critics have used the label "selective outcome reporting." Recent evidence suggests that selective outcome reporting, which tends to produce overestimates of the intervention effects, may be widespread [4,10–13].

For example, a systematic review of the effects of testosterone on erection satisfaction in men with low testosterone identified four eligible trials [14]. The largest trial's results were reported only as "not significant" and could not, therefore, contribute to the meta-analysis. Data from the three smaller trials suggested a large treatment effect (1.3 standard deviations, 95% confidence interval 0.2, 2.3). The review authors ultimately obtained the complete data from the larger trial: after including the less impressive results of the large trial, the magnitude of the effect was smaller and no longer statistically significant (0.8 standard deviations, 95% confidence interval −0.05, 1.63) [15].

The Cochrane handbook suggests that definitive evidence that selective reporting has not occurred requires access to a protocol developed before the study was undertaken [2]. Selective reporting is present if authors acknowledge prespecified outcomes that they fail to report or report outcomes incompletely such that they cannot be included in a meta-analysis. One should suspect reporting bias if the study report fails to include results for a key outcome that one would expect to see in such a study or if composite outcomes are presented without the individual component outcomes.

Note that within the GRADE framework, which rates the quality of a body of evidence, suspicion of selective reporting bias in a number of included studies may lead to rating down of quality of the body of evidence. For instance, in the testosterone example above, had the authors not obtained the missing data, they would have considered rating down the body of evidence for the selective reporting bias suspected in the largest study.

6. Loss to follow-up

Historically, methodologists have sometimes suggested arbitrary thresholds for acceptable loss to follow-up (e.g., less than 20%). The significance of particular rates of loss to follow-up, however, varies widely and is dependent on the relation between loss to follow-up and number of events. For instance, loss to follow-up of 5% in both intervention and control groups would entail little threat of bias if event rates were 20% and 40% in intervention and control groups, respectively. If event rates were 2% and 4%, however, concern with 5% loss to follow-up is much greater.

To state this as a general rule, the higher the proportion lost to follow-up in relation to intervention and control event rates, and differences between intervention and control groups, the greater the threat of bias. Even with relatively high rates of loss to follow-up, however, bias will result only if the number lost is imbalanced between groups or the relationship between loss to follow-up and the likelihood of events differs between intervention and control groups. Unfortunately, we never know if the relationship between loss to follow-up and the likelihood of events does or does not differ in intervention and control groups; large loss to follow-up in relation to the number of events always, therefore, raises the issue of a serious threat of bias.

The issue is conceptually identical with continuous outcomes: Was the loss to follow-up such that reasonable assumptions about differences in outcomes among those lost to follow-up in intervention and control groups could change the overall results in an important way? One can test a variety of assumptions about rates of events in those lost to follow-up when the outcome is a binary variable. One can also conduct such sensitivity analyses when the data are continuous, although the statistical modeling is more challenging.

7. Study limitations in observational studies

Systematic reviews of tools to assess the methodological quality of nonrandomized studies have identified more than 200 checklists and instruments [16–19]. Table 2 summarizes key criteria for observational studies that reflect the contents of these checklists. Judgments associated with assessing study limitations in observational studies are often complex; here, we address two key issues that arise in assessing risk of bias.

7.1. Case series: the problem of missing internal controls

Ideally, observational studies will choose contemporaneous comparison groups that, as far as possible, differ from intervention groups only in the decision (typically by

Table 2
Study limitations in observational studies

1. Failure to develop and apply appropriate eligibility criteria (inclusion of control population)
Under- or overmatching in case–control studies
Selection of exposed and unexposed in cohort studies from different populations
2. Flawed measurement of both exposure and outcome
Differences in measurement of exposure (e.g., recall bias in case–control studies)
Differential surveillance for outcome in exposed and unexposed in cohort studies
3. Failure to adequately control confounding
Failure of accurate measurement of all known prognostic factors
Failure to match for prognostic factors and/or lack of adjustment in statistical analysis
4. Incomplete follow-up

patient or clinician) not to use the intervention. Researchers will enroll and observe intervention and comparison group patients in identical ways. This is the prototypical design using what might be called ''internal controls''—internal, that is, to the study under conduct.

An alternative approach is to study only patients exposed to the intervention—a design we refer to as a case series (others may use ''single group cohort''). To make inferences regarding intervention effects, case series must still refer to results in a comparison group. In many case series, however, the source of comparison group results is implicit or unclear. Such vagueness raises serious questions about the prognostic similarity of intervention and comparison groups and will usually warrant rating down from low- to very low-quality evidence. For instance, in considering the relative impact of low–molecular weight heparin vs. unfractionated heparin in pregnant women, we find systematic reviews of the incidence of bleeding in women receiving the former agent [20,21] but no direct comparisons with the latter.

Thus, case series typically yield very low-quality evidence. There are, however, exceptions. Consider the question of the impact of routine colonoscopy vs. no screening for colon cancer on the rate of perforation associated with colonoscopy. Here, a large series of representative patients undergoing colonoscopy will provide high-quality evidence. When control rates are near zero, case series of representative patients (one might call these cohort studies) can provide high-quality evidence of adverse effects associated with an intervention. One should not confuse these with isolated case reports of associations between exposures and rare adverse outcomes (as have, for instance, been reported with vaccine exposure).

7.2. *Dealing with prognostic imbalance*

Observational studies are at risk of bias because of differences in prognosis in exposed and unexposed populations; to the extent that the two groups come from the same time, place, and population, this risk of bias is diminished. Nevertheless, prognostic imbalance threatens the validity of all observational studies. If the available studies have failed to measure known important prognostic factors, have measured them badly, or have failed to take these factors into account in their analysis (by matching or statistical adjustment), review authors and guideline developers should consider rating down the quality of the evidence from low to very low.

For example, a cohort study using a large administrative database demonstrated an increased risk of cancer-related mortality in diabetic patients using sulfonylureas or insulin relative to metformin [22]. The investigators did not have data available and could, therefore, not adjust for key prognostic variables, including smoking, family history of cancer, occupational exposure, dietary history, and exposure to pollutants. Thus, the study—and others like it that fail to adjust for key prognostic variables—provides only very low-quality evidence of a causal relation between the hypoglycemic agent and cancer deaths.

8. Limitations of GRADE's approach to assessing risk of bias in individual studies

GRADE's approach to assessing risk of bias shares two fundamental limitations with the very large number of alternative approaches. First, empirical evidence supporting the criteria is limited—attempts to show systematic difference between studies that meet and do not meet specific criteria have shown inconsistent results. Second, the relative weight one should put on the criteria remains uncertain.

The GRADE approach is less comprehensive than many systems, emphasizing simplicity and parsimony over completeness. GRADE's approach does not provide a quantitative rating of risk of bias. Although such a rating has advantages, we share with the Cochrane Collaboration methodologists a reluctance to provide a risk of bias score that, by its nature, must make questionable assumptions about the relative extent of bias associated with individual items and fails to consider the context of the individual items.

9. Summarizing study limitations must be outcome specific

Sources of bias may vary in importance across outcomes. Thus, within a single study, one may have higher quality evidence for one outcome than for another. For instance, RCTs of steroids for acute spinal cord injury measured both all-cause mortality and, based on a detailed physical examination, motor function [23–25]. Blinding of outcome assessors is irrelevant for mortality but crucial for motor function. Thus, as in this example, if the outcome assessors in the primary studies on which a guideline panel relies were not blinded, the panel might categorize evidence for all-cause mortality as having no serious study limitations and rate down the evidence for motor function by one level on the basis of serious study limitations.

10. Summarizing risk of bias requires consideration of all relevant evidence

Every study addressing a particular outcome will differ, to some degree, in risk of bias. Review authors and guideline developers must make an overall judgment, considering all the evidence, whether quality of evidence for an outcome warrants rating down on the basis of study limitations.

Table 3 presents the structure of GRADE's approach to study limitations in RCTs. The second column in Table 3 presents the approach as applied to individual studies; the remaining columns refer to the entire body of evidence. Individual trials achieve a low risk of bias when most or all key criteria are met and any violations are not crucial. Studies that suffer from one crucial violation—a violation of crucial importance with regard to a point estimate (in the

Table 3
Summarizing study limitations for randomized trials

Extent of risk of bias	Risk of bias within a study	Risk of bias across studies	Interpretation across studies[a]	Example of summary across studies
No serious limitations, do not downgrade	Low risk of bias for all key criteria (Table 1)	Most information is from studies at low risk of bias	High-quality evidence: the true effect lies close to that of the estimate of the effect	Beta-blockers reduce mortality in patients with heart failure [26]
Serious limitations, rate down one level (i.e., from high to moderate quality)	Crucial limitation for one criterion or some limitations for multiple criteria sufficient to lower ones confidence in the estimate of effect	Most information is from studies at moderate risk of bias	Quality of evidence reduced from high- to moderate-quality evidence: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different	Amodiaquine and SP together likely reduce treatment failures compared with SP alone in patients with malaria [27]
Very serious limitations rate down two levels (i.e., from high to low quality or moderate to very low)	Crucial limitation for one or more criteria sufficient to substantially lower ones confidence in the estimate of effect	Most information is from studies at high risk of bias	Quality of evidence reduced from high- to low-quality evidence: the true effect may be substantially different from the estimate of the effect	Open discectomy may reduce symptoms after 1 yr compared with conservative treatment of lumbar disc prolapse [28]

Abbreviation: SP, sulfadoxine-pyrimethamine.
[a] This interpretation assumes no problems that necessitate rating down because of imprecision, inconsistency, indirectness, and publication bias.

context of a systematic review) or decision (in the context of a guideline)—provide limited-quality evidence. When one or more crucial limitations substantially lower confidence in a point estimate, a body of evidence provides only very limited support for inferences regarding the magnitude of a treatment effect.

Table 3 illustrates that high-quality evidence is available when most studies from a body of evidence meet bias-minimizing criteria. For example, of the 22 trials addressing the impact of beta-blockers on mortality in patients with heart failure, most, probably or certainly, used concealed allocation, all blinded at least some key groups, and follow up of randomized patients was almost complete [26].

GRADE considers a body of evidence of moderate quality when the best evidence comes from individual studies of moderate quality. For instance, we cannot be confident that, in patients with falciparum malaria, amodiaquine and sulfadoxine-pyrimethamine together reduce treatment failures compared with sulfadoxine-pyrimethamine alone because the apparent advantage of sulfadoxine-pyrimethamine was sensitive to assumptions regarding the event rate in those lost to follow-up in two of three studies [27].

Surgery vs. conservative treatment in the management of patients with lumbar disc prolapse provides an example of rating down two levels because of risk of bias in RCTs [28]. We are uncertain of the benefit of open disectomy in reducing symptoms after 1 year or longer because of very serious limitations in one trial of open disectomy compared with conservative treatment without a large number of early crossovers in both comparison groups. That trial suffered from inadequate concealment of allocation and unblinded assessment of outcome by potentially biased raters (surgeons) using unvalidated rating instruments (Table 4).

11. Existing systematic reviews are often limited in summarizing study limitations across studies

To rate overall quality of evidence with respect to an outcome, review authors and guideline developers must

Table 4
Quality assessment for open discectomy vs. conservative treatment (Gibson and Waddell [28])

Quality assessment						
No of patients (studies)	Design	Limitations	Inconsistency	Indirectness	Imprecision	Publication bias
Outcome: poor/bad result at 1yr—surgeon rated						
126 (1)	RCT	Very serious limitations[a]	Not relevant	No serious indirectness	Serious imprecision[b]	Unlikely
Outcome: poor/bad result at 4yr—surgeon rated						
126 (1)	RCT	Very serious limitations[a]	Not relevant	No serious indirectness	Serious imprecision[b]	Unlikely
Outcome: poor/bad result at 10yr—surgeon rated						
126 (1)	RCT	Very serious limitations[a]	Not relevant	No serious indirectness	Serious imprecision[b]	Unlikely

Abbreviation: RCT, randomized controlled trial.
[a] Inadequate concealment of allocation and unblinded unvalidated assessment by the surgeon.
[b] Wide confidence intervals and few events (16 or fewer).

consider and summarize study limitations considering all the evidence from multiple studies. For a guideline developer, using an existing systematic review would be the most efficient way to address this issue.

Unfortunately, systematic reviews usually do not address all important outcomes, typically focusing on benefit and neglecting harm. For instance, one is required to go to separate reviews to assess the impact of beta-blockers on mortality [26] and on quality of life [29]. No systematic review has addressed beta-blocker toxicity in heart failure patients.

Review authors' usual practice of rating the quality of studies across outcomes, rather than separately for each outcome, further limits the usefulness of existing systematic reviews for guideline developers. This approach becomes even more problematic when review authors use summary measures that aggregate across quality criteria (e.g., allocation concealment, blinding, loss to follow-up) to provide a single score. These measures are often limited in that they focus on quality of reporting rather than on the design and conduct of the study [30]. Furthermore, they tend to be unreliable and less closely correlated with outcome than individual quality components [31−33]. These problems arise, at least in part, because calculating a summary score inevitably involves assigning arbitrary weights to different criteria.

Finally, systematic reviews that address individual components of study limitations are often not comprehensive and fail to make transparent the judgments needed to evaluate study limitations. These judgments are often challenging, at least in part, because of inadequate reporting: just because a safeguard against bias is not reported does not mean it was neglected [34,35].

Thus, although systematic reviews are often extremely useful in identifying the relevant primary studies, members of guideline panels or their delegates must often review individual studies if they wish to ensure accurate ratings of study limitations for all relevant outcomes. As review authors increasingly adopt the GRADE approach (and in particular as Cochrane review authors do so in combination with using the Cochrane risk of bias tool), the situation will improve.

12. What to do when there is only one RCT

Many people are uncomfortable designating a single RCT as high-quality evidence. Given the many instances in which the first positive report has not held up under subsequent investigation, this discomfort is warranted. On the other hand, automatically rating down quality when there is a single study is not appropriate. A single, very large, rigorously planned and conducted multicentre RCT may provide high-quality evidence. GRADE suggests especially careful scrutiny of all relevant issues (risk of bias, precision, directness, and publication bias) when only a single RCT addresses a particular question.

13. Moving from Cochrane risk of bias tables in individual studies to rating quality of evidence across studies

Moving from 6 risk of bias criteria for each individual study to a judgment about rating down for quality of evidence for risk of bias across a group of studies addressing a particular outcome presents challenges. We suggest the following principles.

First, in deciding on the overall quality of evidence, one does not average across studies (for instance if some studies have no serious limitations, some serious limitations, and some very serious limitations, one does not automatically rate quality down by one level because of an average rating of serious limitations). Rather, judicious consideration of the contribution of each study, with a general guide to focus on the high-quality studies (as we will illustrate), is warranted.

Second, this judicious consideration requires evaluating the extent to which each trial contributes toward the estimate of magnitude of effect. This contribution will usually reflect study sample size and number of outcome events—larger trials with many events will contribute more, much larger trials with many more events will contribute much more.

Third, one should be conservative in the judgment of rating down. That is, one should be confident that there is substantial risk of bias across most of the body of available evidence before one rates down for risk of bias.

Fourth, the risk of bias should be considered in the context of other limitations. If, for instance, reviewers find themselves in a close-call situation with respect to two quality issues (risk of bias and, say, precision), we suggest rating down for at least one of the two.

Fifth, notwithstanding the first four principles, reviewers will face close-call situations. They should both acknowledge that they are in such a situation, make it explicit why they think this is the case, and make the reasons for their ultimate judgment apparent.

14. Application of principles

A systematic review of flavonoids to treat pain and bleeding associated with hemorrhoids [36], with respect to the primary outcome of persisting symptoms, most trials did not provide sufficient information to determine whether randomization was concealed, the majority violated the intention-to-treat principle and did not provide the data allowing the appropriate analysis (Table 5), and none used a validated symptom measure. On the other hand, most authors described their trials as double blind, and although concealment and blinding are different concepts, blinded trials of drugs are very likely to be concealed [34] (Table 5). Because the questionnaires appeared simple and transparent, and because of the blinding of the studies, we would be hesitant to consider lack of validation introducing a serious risk of bias.

Table 5
Risk of bias for measurement of symptoms in studies of flavonoids in patients with hemorrhoids

Study[c]	Randomization	Allocation concealment	Blinding	Loss to follow-up[a]/IT principle observed or per protocol analysis	Other
Dimitroulopoulos D, 2005	Adequate[b] Computer-generated random numbers[b]	Sealed opaque envelopes[b]	Described as single blind Care givers, patients, and data collectors blinded[b]	3%/protocol	Unvalidated symptom measure
Misra MC, 2000	Adequate Computer-generated random numbers[b]	Adequate Sealed opaque envelopes[b]	Patients and physicians[b] Described as double blind Placebo identical appearance	2%/protocol	Unvalidated symptom measure
Godeberge P, 1994	Adequate[b]	Adequate Sealed opaque envelopes[b]	Patients, physician-investigator, data manager, statistician, and authors blinded	6%/protocol	
Cospite M, 1994	Unclear	Unclear	Unclear Described as double blind	12%/IT	Unvalidated symptom measure
Chauvenet-M, 1994	Unclear	Unclear	Unclear	11%/protocol	Unvalidated symptom measure
Ho Y-H, 2000	Adequate Drawing of sealed opaque envelopes[b]	Adequate Sealed opaque envelopes	All parties blinded[b]	0%/IT	Unvalidated symptom measure
Thanapongsathorn W, 1992	Unclear	Unclear	Unclear Described as double blind	I2%/protocol	Unvalidated symptom measure
Titapant V, 2001	Unclear	Unclear	Unclear Described as double blind Placebo identical appearance	12%/protocol	Unvalidated symptom measure
Wijayanegara H, 1992	Unclear	Unclear	Unclear Described as double blind	3%/protocol	Unvalidated symptom measure
Annoni F, 1986	Unclear	Unclear	Unclear Described as double blind Placebo identical appearance	Uncertain/unclear	Unvalidated symptom measure
Thorp RH, 1970	Unclear	Unclear	Physicians and patients blinded Described as double blind Placebo identical appearance	20%/protocol	Unvalidated symptom measure
Clyne MB, 1967	Bottles numbered consecutively in accordance to random tables	Unclear	Physicians and patients blinded Described as double blind Placebo identical appearance	Uncertain/protocol	Unvalidated symptom measure
Sinnatamby CS, 1973	Unclear	Unclear	Physicians and patients blinded Described as double blind	53%/protocol	Unvalidated symptom measure
Trochet JP, 1992	Randomized by blocks of three (method unclear)	Unclear	Physicians blinded Placebo identical appearance	Uncertain/IT	Unvalidated symptom measure

Abbreviation: IT, intention-to-treat principle observed.
[a] No important differences in rate of loss to follow-up between flavonoid and control groups in any study.
[b] Data provided by authors.
[c] For full citation of the references cited in this table, see Alonso-Coello et al.[36]

Nevertheless, in light of these study limitations, one might consider focusing on the highest quality trials. Substantial precision would, however, be lost (requiring rating down for imprecision), and the quality of the trials did not explain variability in results (i.e., the magnitude of effect was similar in the methodologically stronger and weaker studies). Both considerations argue for basing an estimate on the results of all RCTs.

In our view, this represents a borderline situation in which it would be reasonable either to rate down for risk of bias or not to do so. This illustrates that the great merit of GRADE is not that it ensures consistency of conclusions but that it requires explicit and transparent judgments. Considering these issues in isolation, and following the principles articulated above, however, we would be inclined not to rate down for quality for risk of bias.

The possibility of discrepant judgments between intelligent and well-informed review authors is more than theoretical. A number of RCTs have evaluated the extent to which graduated pressure stockings can prevent deep venous thrombosis (DVT) in airline passengers taking long flights. Cochrane review authors concluded that the studies provided high-quality evidence for DVT prevention [37]. In contrast, a group of thrombosis experts involved in producing a guideline concluded that because of use of an unreliable method of diagnosing DVT, and lack of blinding, the evidence was of

low quality [38]. Even after direct contact and discussion, each group adhered to its own position—and it remains possible that either group is correct.

Three RCTs addressing the impact of 24-hour administration of high-dose corticosteroids on motor function in patients with acute spinal cord injury illustrate another principle of aggregation [23–25]. Although the degree of limitations is in fact a continuum (as Fig. 1 illustrates), GRADE simplifies the process by categorizing these studies—or any other study—as having "no serious limitations," "serious limitations," or "very serious limitations" (as in Table 3).

The first of the three trials (Bracken in Fig. 1), which included 127 patients treated within 8 hours of injury, ensured allocation concealment through central randomization, almost certainly blinded patients, clinicians, and those measuring motor function, and lost 5% of patients to follow-up at 1 year [23]. The flaws in this RCT are sufficiently minor to allow classification as "no serious limitations."

The second trial (Pointillart et al. [25] in Fig. 1) was unlikely to have concealed allocation, did blind those assessing outcome (but not patients or clinicians), and lost only one of 106 patients to follow-up. Here, quality falls in an intermediate range, and classification as either "no serious limitations" or "serious limitations" may be appropriate. The third trial (Otani et al. [24] in Fig. 1), which included 158 patients, almost certainly failed to conceal allocation, used no blinding, and lost 26% of patients to follow-up, many more in the steroid group than the control group. This third trial is probably best classified as having "very serious limitations."

Considering these three RCTs, should one rate down for design and implementation with respect to the motor function outcome? If we considered only the first two trials, the answer would be no. Therefore, the review authors must decide either to exclude the third trial (thereby only including trials with few limitations) or include it based on a judgment that overall there is a low risk of bias (because most of the evidence comes from trials with few limitations) despite the contribution of the trial with very serious limitations to the overall estimate of effect. This example illustrates that averaging across studies will not be the right approach.

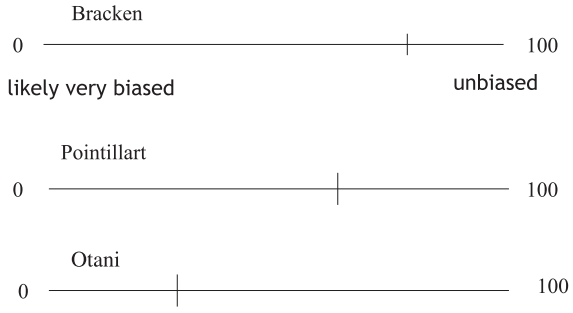

Fig. 1. Validity of three randomized controlled trials addressing the effect of steroids on motor function in acute spinal cord injury.

15. Recording judgments about study limitations

One great merit of GRADE is its lucid categorization of factors that decrease quality of evidence and the resultant transparency of judgments. This transparency, however, requires careful documentation of judgments. Including a risk of bias table that summarizes key criteria used to assess study limitations for each outcome for each study helps ensure transparency.

Table 5 presents an example of such a table. Note that the table focuses on only one outcome, symptoms. Each study will need only one line on such a table if, as in this case, there is only one important outcome or if each quality criterion is the same for every important outcome. Each outcome for which quality criteria differ in important ways will need a separate line. Outcomes may, for instance, differ for blinding (e.g., in surgical trials patients completing questionnaires measuring health-related quality of life may be unblinded, but adjudicators of cause-specific mortality may be blinded) or loss to follow-up (e.g., greater loss to follow-up for quality of life than for all-cause mortality).

Review authors and guideline developers can then summarize their assessments across studies in a "quality assessment" table to fully ensure the transparency of their judgments (Table 4). A footnote provides the reasoning behind the decision to rate down the quality of the evidence from high to low quality on the basis of study limitations (alternatively, one can very briefly summarize the key information in a cell in the table). In this example, there was an additional concern about imprecision, which further decreases the quality of evidence from low to very low. We will describe guidelines for making judgments about imprecision (the risk of random error), in the sixth article in this series.

References

[1] The users' guides to the medical literature: a manual for evidence-based clinical practice. In: Guyatt G, Rennie D, Meade M, Cook D, editors. 2nd ed. New York, NY: McGraw-Hill; 2008.
[2] Higgins JP, Altman D. Assessing the risk of bias in included studies. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions 5.0.1. Chichester, UK: John Wiley & Sons; 2008.
[3] Vedula SS, Bero L, Scherer RW, Dickersin K. Outcome reporting in industry-sponsored trials of gabapentin for off-label use. N Engl J Med 2009;361:1963–71.
[4] Mathieu S, Boutron I, Moher D, Altman DG, Ravaud P. Comparison of registered and published primary outcomes in randomized controlled trials. JAMA 2009;302:977–84.
[5] Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336:601–5.
[6] Pocock SJ. When (not) to stop a clinical trial for benefit. JAMA 2005;294:2228–30.
[7] Pocock SJ, Hughes MD. Practical problems in interim analyses, with particular regard to estimation. Control Clin Trials 1989;10(4 Suppl): 209S–21S.

[8] Montori VM, Devereaux PJ, Adhikari NK, Burns KE, Eggert CH, Briel M, et al. Randomized trials stopped early for benefit: a systematic review. JAMA 2005;294:2203−9.
[9] Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. JAMA 2010;303:1180−7.
[10] Furukawa TA, Watanabe N, Omori IM, Montori VM, Guyatt GH. Association between unreported outcomes and effect size estimates in Cochrane meta-analyses. JAMA 2007;297:468−70.
[11] Chan AW, Altman DG. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. BMJ 2005;330:753.
[12] Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291:2457−65.
[13] Chan AW, Krleza-Jerić K, Schmid I, Altman DG. Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. CMAJ 2004;171:735−40.
[14] Bolona ER, Uraga MV, Haddad RM, Tracz MJ, Sideras K, Kennedy CC, et al. Testosterone use in men with sexual dysfunction: a systematic review and meta-analysis of randomized placebo-controlled trials. Mayo Clin Proc 2007;82:20−8.
[15] Sinha M, Montori VM. Reporting bias and other biases affecting systematic reviews and meta-analyses: a methodological commentary. Expert Rev Pharmacoecon Outcomes Res 2006;6(1):603−11.
[16] Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:iii−x. 1−173.
[17] West S, King V, Carey TS, Lohr KN, Mckoy N, Sutton SF, et al. Systems to rate the strength of scientific evidence [evidence report/technology assessment no 47]. AHRQ Publication No 02−E016. Rockville, MD: Agency for Healthcare Research and Quality; 2002.
[18] Proposed Evaluation Tools for COMPUS: assessment, November 29, 2005. Ottawa, Canada: Canadian Coordinating Office for Health Technology; 2005.
[19] Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. Int J Epidemiol 2007;36:666−76.
[20] Greer IA, Nelson-Piercy C. Low-molecular-weight heparins for thromboprophylaxis and treatment of venous thromboembolism in pregnancy: a systematic review of safety and efficacy. Blood 2005;106:401−7.
[21] Sanson BJ, Lensing AW, Prins MH, Ginsberg JS, Barkagan ZS, Lavenne-Pardonge E, et al. Safety of low-molecular-weight heparin in pregnancy: a systematic review. Thromb Haemost 1999;81:668−72.
[22] Bowker SL, Majumdar SR, Veugelers P, Johnson JA. Increased cancer-related mortality for patients with type 2 diabetes who use sulfonylureas or insulin. Diabetes Care 2006;29:254−8.
[23] Bracken MB, Shepard MJ, Collins WF Jr, Holford TR, Baskin DS, Eisenberg HM, et al. Methylprednisolone or naloxone treatment after acute spinal cord injury: 1-year follow-up data. Results of the second National Acute Spinal Cord Injury Study. J Neurosurg 1992;76:23−31.
[24] Otani K, Abe H, Kadoya S. Beneficial effect of methylprednisolone sodium succinate in the treatment of acute spinal cord injury. Sekitsui Sekizui 1994;7:633−47.
[25] Pointillart V, Petitjean ME, Wiart L, Vital JM, Lassié P, Thicoipé M, et al. Pharmacological therapy of spinal cord injury during the acute phase. Spinal Cord 2000;38:71−6.
[26] Brophy JM, Joseph L, Rouleau JL. Beta-blockers in congestive heart failure. A Bayesian meta-analysis. Ann Intern Med 2001;134:550−60.
[27] McIntosh H, Jones K. Chloroquine or amodiaquine combined with sulfadoxine-pyrimethamine for treating uncomplicated malaria. Cochrane Database Syst Rev 2005;(4). CD000386.
[28] Gibson J, Waddell G. Surgical interventions for lumbar disc prolapse. Cochrane Database Syst Rev 2007;(2)10.1002/14651858. CD001350.
[29] Dobre D, van Jaarsveld CH, deJongste MJ, Haaijer Ruskamp FM, Ranchor AV. The effect of beta-blocker therapy on quality of life in heart failure patients: a systematic review and meta-analysis. Pharmacoepidemiol Drug Saf 2007;16:152−9.
[30] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62−73.
[31] Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408−12.
[32] Emerson JD, Burdick E, Hoaglin DC, Mosteller F, Chalmers TC. An empirical study of the possible relation of treatment differences to quality scores in controlled randomized clinical trials. Control Clin Trials 1990;11:339−52.
[33] Juni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. JAMA 1999;282:1054−60.
[34] Devereaux PJ, Choi PT, El-Dika S, Bhandari M, Montori VM, Schünemann HJ, et al. An observational study found that authors of randomized controlled trials frequently use concealment of randomization and blinding, despite the failure to report these methods. J Clin Epidemiol 2004;57:1232−6.
[35] Soares HP, Daniels S, Kumar A, Clarke M, Scott C, Swann S, Djulbegovic B, et al. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. BMJ 2004;328:22−4.
[36] Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, Mills E, Heels-Ansdell D, Johanson JF, et al. Meta-analysis of flavonoids for the treatment of haemorrhoids. Br J Surg 2006;93(8):909−20.
[37] Clarke M, Hopewell S, Juszczak E, Eisinga A, Kjeldstrøm M. Compression stockings for preventing deep vein thrombosis in airline passengers. Cochrane Database Syst Rev 2007;(3). CD004002.
[38] Geerts WH, Bergqvist D, Pineo GF, Heit JA, Samama CM, Lassen MR, et al. Prevention of venous thromboembolism: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines (8th Edition). Chest 2008;133(6 Suppl):381S−453S.