ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1294–1302

Journal of Clinical Epidemiology

GRADE guidelines: 7. Rating the quality of evidence—inconsistency

Gordon H. Guyatt[a,b,*], Andrew D. Oxman[c], Regina Kunz[d], James Woodcock[e], Jan Brozek[a], Mark Helfand[f], Pablo Alonso-Coello[g], Paul Glasziou[h], Roman Jaeschke[b], Elie A. Akl[i], Susan Norris[j], Gunn Vist[c], Philipp Dahm[k], Vijay K. Shukla[l], Julian Higgins[m], Yngve Falck-Ytter[n], Holger J. Schünemann[a,b], The GRADE Working Group[1]

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada*
[b]*Department of Medicine, McMaster University, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada*
[c]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, 0130 Oslo, Norway*
[d]*Basel Institute of Clinical Epidemiology, University Hospital Basel Hebelstrasse 10, 4031 Basel, Switzerland*
[e]*London School of Hygiene and Tropical Medicine, London, United Kingdom*
[f]*Oregon Evidence-Based Practice Center, Oregon Health & Science University, Portland VA Medical Center, Portland, OR, USA*
[g]*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*
[h]*Centre for Research in Evidence-Based Practice, Faculty of Health Sciences, Bond University, Gold Coast, Queensland, 4229, Australia*
[i]*Department of Medicine, State University of New York at Buffalo, NY, USA*
[j]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, USA*
[k]*Department of Urology, University of Florida, College of Medicine, Gainesville, FL 3210, USA*
[l]*Canadian Agency for Drugs and Technology in Health (CADTH), Ottawa K1S 5S8, Canada*
[m]*MRC Biostatistics Unit, Cambridge, United Kingdom*
[n]*Division of Gastroenterology, Case Medical Center and VA, Case Western Reserve University, Cleveland, OH 44106, USA*

Accepted 8 March 2011; Published online 31 July 2011

Abstract

This article deals with inconsistency of relative (rather than absolute) treatment effects in binary/dichotomous outcomes. A body of evidence is not rated up in quality if studies yield consistent results, but may be rated down in quality if inconsistent. Criteria for evaluating consistency include similarity of point estimates, extent of overlap of confidence intervals, and statistical criteria including tests of heterogeneity and I^2. To explore heterogeneity, systematic review authors should generate and test a small number of a priori hypotheses related to patients, interventions, outcomes, and methodology. When inconsistency is large and unexplained, rating down quality for inconsistency is appropriate, particularly if some studies suggest substantial benefit, and others no effect or harm (rather than only large vs. small effects).

Apparent subgroup effects may be spurious. Credibility is increased if subgroup effects are based on a small number of a priori hypotheses with a specified direction; subgroup comparisons come from within rather than between studies; tests of interaction generate low *P*-values; and have a biological rationale.

Keywords: GRADE; Inconsistency; Heterogeneity; Variability; Sub-group analysis; Interaction

1. Introduction

Previous articles in this series presenting GRADE's approach to systematic reviews and clinical guidelines have dealt with framing the question, defined quality of evidence, and described GRADE's approach to rating down the quality of a body of evidence because of problems with bias and imprecision. This article deals with inconsistency in the magnitude of effect in studies of alternative management strategies; it does not address inconsistency in diagnostic test studies.

[1] 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's Web site at www.elsevier.com.

* Corresponding author. CLARITY Research Group, Department of Clinical Epidemiology and Biostatistics, McMaster University, Room 2C12, 1200 Main Street, West Hamilton, Ontario L8N 3Z5, Canada. 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.2011.03.017

Key points

- GRADE suggests rating down the quality of evidence if large inconsistency (heterogeneity) in study results remains after exploration of a priori hypotheses that might explain heterogeneity.
- Judgment of the extent of heterogeneity is based on similarity of point estimates, extent of overlap of confidence intervals, and statistical criteria including tests of heterogeneity and I^2.
- Apparent subgroup effects should be interpreted cautiously with attention to whether subgroup comparisons come from within rather than between studies; if tests of interaction generate low *P*-values; and whether subgroup effects are based on a small number of a priori hypotheses with a specified direction.

1.1. This article deals with binary/dichotomous outcomes, and inconsistency in relative, not absolute, measures of effect

Patients vary widely in their preintervention or baseline risk of the adverse outcomes that health care interventions are designed to prevent (e.g., death, stroke, myocardial infarction, disease exacerbation). As a result, risk differences (absolute risk reductions) in subpopulations tend to vary widely. Relative risk (RR) reductions, on the other hand, tend to be similar across subgroups, even if subgroups have substantial differences in baseline risk [1−3]. Therefore, when we refer to inconsistencies in effect size, we are referring to relative measures (risk ratios and hazard ratios—which we prefer—or odds ratios).

GRADE considers the issue of difference in absolute effect in subgroups of patients—much more common than differences in relative effect—as a separate issue. When easily identifiable patient characteristics confidently permit classifying patients into subpopulations at appreciably different risk, absolute differences in outcome between intervention and control groups will differ substantially between these subpopulations. This may well warrant differences in recommendations across subpopulations. We deal with the issue of subpopulations whose baseline risk differs in other articles in this series [4,5].

1.2. We rate down for inconsistency, not up for consistency

We pointed out in a previous article in this series [6] that consistent results do not mandate rating up quality of evidence. The reason is that a consistent bias will lead to consistent, spurious findings. Such consistent biases are often plausible (health-conscious individuals make consistently different decisions than those who are less health conscious; a variety of factors lead to consistently better health in high vs. low socioeconomic status individuals).

1.3. Large inconsistency demands a search for an explanation

Systematic review authors should be prepared to face inconsistency in the results. In the early (protocol) stages of their review, they should consider the diversity of patients, interventions, outcomes that may be appropriate to include. Reviewers should combine results only if, across the range of patients, interventions, and outcomes considered, it is plausible that the underlying magnitude of treatment effect is similar [7]. This decision is a matter of judgment. In general, we suggest beginning by pooling widely, and then testing whether the assumption of similar effects across studies holds. This approach necessitates generating a priori hypotheses regarding possible explanations of variability of results.

If systematic review authors find that the magnitude of intervention effects differs across studies, explanations may lie in the population (e.g., disease severity), the interventions (e.g., doses, cointerventions, comparison interventions), the outcomes (e.g., duration of follow-up), or the study methods (e.g., randomized trials with higher and lower risk of bias). If one of the first three categories provides the explanation, review authors should offer different estimates across patient groups, interventions, or outcomes. Guideline panelists are then likely to offer different recommendations for different patient groups and interventions. If study methods provide a compelling explanation for differences in results between studies, then authors should consider focusing on effect estimates from studies with a lower risk of bias.

If large variability (often referred to as heterogeneity) in magnitude of effect remains unexplained, the quality of evidence decreases. In this article, we provide guidance concerning how to judge whether inconsistency in results is sufficient to rate down the quality of evidence, and when to believe apparent explanations of inconsistency (subgroup analyses).

1.4. Four criteria for assessing inconsistency in results

Reviewers should consider rating down for inconsistency when

1. Point estimates vary widely across studies;
2. Confidence intervals (CIs) show minimal or no overlap;
3. The statistical test for heterogeneity—which tests the null hypothesis that all studies in a meta-analysis have the same underlying magnitude of effect—shows a low *P*-value;
4. The I^2—which quantifies the proportion of the variation in point estimates due to among-study differences—is large.

One may ask: what is a large I^2? One set of criteria would say that an I^2 of less than 40% is low, 30–60% may be moderate, 50–90% may be substantial, and 75–100% is considerable [8]. Note the overlapping ranges, and the equivocation (''may be''): an implicit acknowledgment that the thresholds are both arbitrary and uncertain.

Furthermore, although it does not—in contrast to tests for heterogeneity—depend on the number of studies, I^2 shares limitations traditionally associated with tests for heterogeneity. When individual study sample sizes are small, point estimates may vary substantially but, because variation may be explained by chance, I^2 may be low. Conversely, when study sample size is large, a relatively small difference in point estimates can yield a large I^2 [9]. Another statistic, τ^2 (tau square) is a measure of the variability that has an advantage over other measures in that it is not dependent on sample size [9]. So far, however, it has not seen much use. All statistical approaches have limitations, and their results should be seen in the context of a subjective examination of the variability in point estimates and the overlap in CIs.

1.5. The impact of direction of effect on decisions regarding inconsistency

Consider Fig. 1, a forest plot with four studies, two on either side of the line of no effect. We would have no inclination to rate down for inconsistency. Differences in direction, in and of themselves, do not constitute a criterion for variability in effect if the magnitude of the differences in point estimates is small.

As we define quality of evidence for a guideline, inconsistency is important only when it reduces confidence in results in relation to a particular decision. Even when inconsistency is large, it may not reduce confidence in results regarding a particular decision. Consider, for instance, Fig. 2 in which variability is substantial, but the differences are between small and large treatment effects. Guideline developers may or may not consider this degree of variability important. Because they are, much less than the guideline developers, in a position to judge whether the apparent high heterogeneity can be dismissed on the grounds that it is unimportant, systematic review authors are more likely to rate down for inconsistency. This issue arises in one of the examples—flavonoids in hemorrhoids—that we present subsequently.

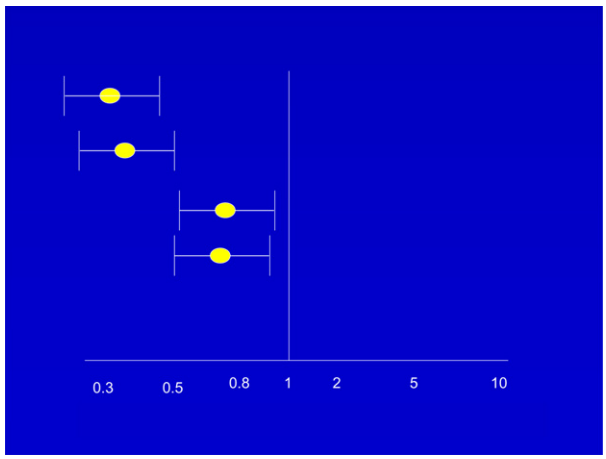

Fig. 2. Substantial heterogeneity, but of questionable importance.

Consider, in contrast, Fig. 3. The magnitude of the variability in results is identical to that of Fig. 2. Here, however, because two studies suggest benefit and two suggest harm, we would unquestionably choose to rate down the quality of evidence as a result of variability in results.

1.6. Test a priori hypotheses about inconsistency even when inconsistency appears to be small

Review authors sometimes set thresholds for the test for heterogeneity (such as $P = 0.1$) or I^2 (such as $I^2 = 30\%$) to determine whether they will search for explanations for inconsistency. The logic is that if the results are very consistent (test for heterogeneity $P > 0.1$, I^2 less than 30%) there is not enough inconsistency to warrant looking for the explanation.

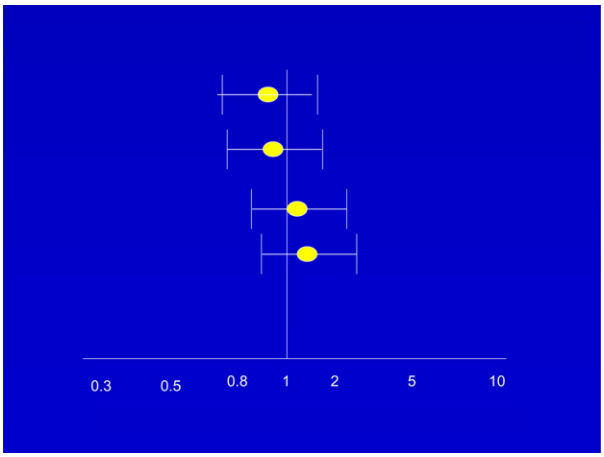

Fig. 1. Differences in direction, but minimal heterogeneity.

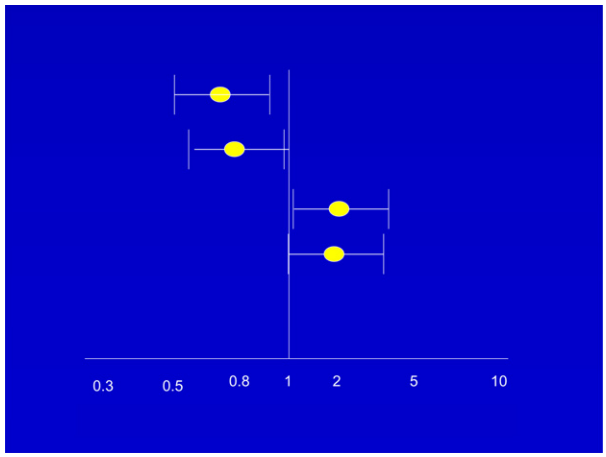

Fig. 3. Substantial heterogeneity, of unequivocal importance.

This is not necessarily the case. For example, a meta-analysis of randomized trials of rofecoxib looking at the outcome of myocardial infarction found apparently consistent results (heterogeneity $P = 0.82$, $I^2 = 0\%$) [10]. Yet, when the investigators examined the effect in trials that used an external endpoint committee (RR 3.88, 95% CI: 1.88, 8.02) vs. trials that did not (RR 0.79, 95% CI: 0.29, 2.13), they found differences that were large and unlikely to be explained by chance ($P = 0.01$).

Although the issue is controversial, we recommend that meta-analyses include formal tests of whether a priori hypotheses explain inconsistency between important subgroups even if the variability that exists appears to be explained by chance (e.g., high *P*-values in tests of heterogeneity, and low I^2 values). As we will discuss below, however, one should always be cautious when interpreting the results of subgroup analyses.

1.7. Rating down for inconsistency: Examples

A systematic review of studies comparing health outcomes in Canada and the United States reported very large differences in effects across studies [11] (Fig. 4). The *P*-value for the test of heterogeneity was <0.0001 and the $I^2 = 94\%$. None of the a priori hypotheses (including study quality, primary data collection vs. administrative database, whether care was primarily outpatient or inpatient, whether the data were collected before or after 1986, and the extent to which US patients had health insurance) explained heterogeneity. Such inconsistency would require rating down by one or (if the quality was not already low because of the observational nature of the studies) two levels (i.e., from high to low, or moderate to very low quality evidence).

A systematic review of flavonoids for symptom relief in patients with hemorrhoids [12] showed wide variation in point estimates and appreciable nonoverlap in CIs, a significant test for heterogeneity ($P = 0.001$) and high I^2 (65.1%) (Fig. 5). The a priori hypotheses (severity and nature of hemorrhoids, cointervention, study quality) failed to explain heterogeneity.

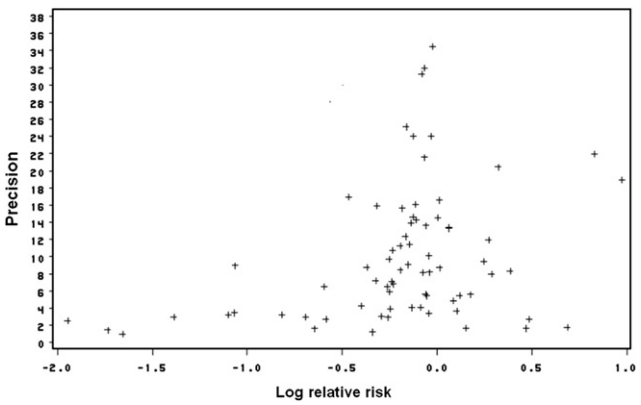

Fig. 4. Funnel plot for all-cause mortality, United States vs. Canada. Negative values favor Canada, positive values, United States.

Despite the inconsistency, the decision to rate down is not straightforward. All studies, with one exception, favor treatment. The inconsistency is therefore almost completely between studies that show moderate, large, and very large effects. Thus, although there is large inconsistency, the importance of the inconsistency for decision making is uncertain. Whether to rate down quality is therefore a matter of judgment.

The argument against rating down for inconsistency in results gains strength from the high control group risk of persisting symptoms (mean value across studies over 56%). Even if the RR reduction is much lower than the pooled estimate of 60%, the risk difference remains substantial (e.g., 20% RR reduction would translate into a risk difference of more than 10 per 100 patients). Thus, the balance of benefits and harms (which are minimal with these agents) is favorable across the range of inconsistent benefits observed. Inconsistency, therefore, has no substantial impact on the judgments required to make a recommendation (so as long as one is confident that there are minimal adverse effects and the cost and bother of taking the medicine is minimal).

1.8. Deciding whether to use estimates from a subgroup analysis

Unexplained inconsistency is undesirable, and resolving the inconsistency far preferable. A satisfactory explanation based on differences in population, interventions, or outcomes mandates generating two (or more) estimates of effect, and tailoring recommendations accordingly. Our examples will come from the most common putative subgroup effect, that related to differences in patients.

Consider, for instance, a systematic review of the use of calcium and vitamin D in preventing osteoporotic fractures in people older than 50 years that suggested a modest 12% reduction in RR of fractures (95% CI: 5, 17) [13]. The effect was minimal in studies focusing on individuals younger than 69 years (RR 0.97), small in those focusing on individuals aged 70−79 years (RR 0.89), and moderate in those focusing on individuals 80 years and older (RR 0.76). If the effect truly differs across subgroups, guideline panels should consider recommending calcium (with or without vitamin D) for the aged, but not for those younger than 69 years.

Unfortunately, there is high likelihood that, in settling on a particular explanation of heterogeneity, one is capitalizing on the play of chance. Indeed, most putative subgroup effects ultimately prove spurious [14]. As a result, reviewers and guideline developers must exercise a high degree of skepticism regarding potential explanations, paying particular attention to criteria in Table 1 [14−16]. Particularly dangerous in the context of conventional (as opposed to individual patient data) meta-analysis is the usual between- rather than within-study nature of the comparison (Table 1). We will illustrate the application of these criteria to three

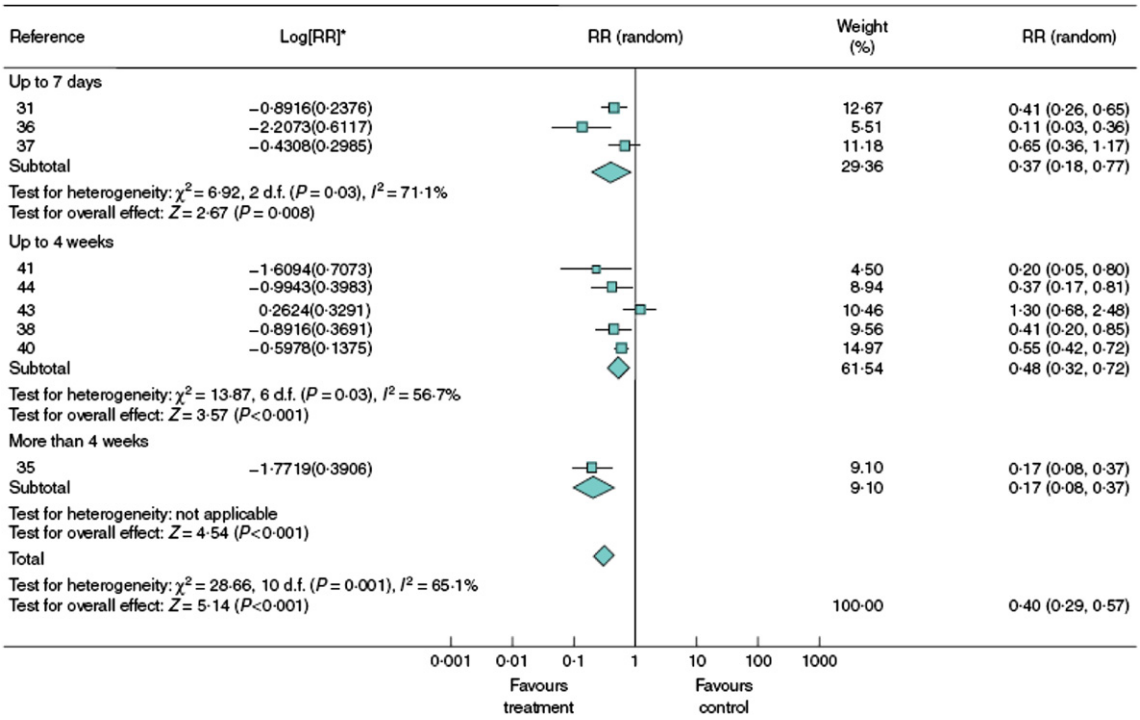

Fig. 5. Results of a systematic review of flavanoids for treatment of hemorrhoids: relative risks of failure to improve.

examples, and the implications for ratings of quality of evidence.

Example 1: A systematic review and individual patient data meta-analysis (IPDMA) addressed the impact of high vs. low positive end-expiratory pressures (PEEPs) in three randomized trials that enrolled 2,299 adult patients with severe acute lung injury requiring mechanical ventilation [17]. IPDMA has two important advantages in elucidating possible subgroup differences. First, all comparisons between subgroups are within study. Secondly, the analysis is much more powerful because it takes advantage of individual patient characteristics rather than summary characteristics of a group of patients included in the study.

The results of this IPDMA suggested a possible reduction in deaths in hospital with the higher PEEP strategy, but the difference was not statistically significant (RR 0.94; 95% CI: 0.86, 1.04). In patients with severe disease (labeled acute respiratory distress syndrome), the effect more clearly favored the high PEEP strategy (RR 0.90; 95% CI: 0.81, 1.00; $P = 0.049$). In patients with mild disease, results suggested that the high PEEP strategy may be inferior (RR 1.37; 95% CI: 0.98, 1.92).

Applying the seven criteria (Table 1), we find that six are met fully, and the seventh, consistency across trials and outcomes, partially: the results of the subgroup analysis were consistent across the three studies, but other ways of measuring severity of lung injury (for instance, treating severity as a continuous variable) failed to show a statistically significant interaction between the severity and the magnitude of effect.

The credibility of subgroup effects is not a matter of yes or no, but a continuum (Fig. 6). In this case, the subgroup analysis is relatively convincing. Therefore, systematic reviewers should present results in both more and less severe patients, and subgroups (as they did) and guideline developers should make recommendations separately for severe and less severe patients.

Example 2: Three randomized trials have tested the effects of vasopressin vs. epinephrine on survival in patients with cardiac arrest [18] (Fig. 7). The results show appreciable differences in point estimates, widely overlapping CIs, a P-value for the test of heterogeneity of 0.21 and an I^2 of 35%.

Two of the trials included both patients in whom asystole was responsible for the cardiac arrest and the patients in whom ventricular fibrillation was the offending rhythm. One of these two trials reported a borderline statistically significant benefit—our own analysis was borderline nonsignificant—of vasopressin over epinephrine restricted to patients with asystole (in contrast to patients whose cardiac arrest was induced by ventricular fibrillation) [19].

Can subgroup analysis of patients with asystole vs. those with ventricular fibrillation explain the moderate inconsistency in the results? Reviewing the seven criteria (Table 1), the answer is ‘‘not very likely.’’ Chance can explain the putative subgroup effect and the hypothesis fails other criteria (including small number of a priori hypotheses and consistency of effect). Here, guideline developers should make recommendations on the basis of the pooled estimate of data from both the groups. Whether the quality of evidence should

Table 1. Criteria for judging the credibility of subgroup analyses with examples

Criterion	Example 1: High vs. low positive end-expiratory pressure (PEEP) in more vs. less severe patients with acute lung injury	Example 2: Vasopressin vs. epinephrine in cardiac arrest: asystole vs. ventricular fibrillation	Example 3: Calcium for fracture prevention in older vs. younger individuals
Is the subgroup variable a characteristic specified at baseline (in contrast with after randomization)?	Yes	Yes	Yes
Is the subgroup difference suggested by comparisons within rather than between studies?	Yes	Two of three within-study comparisons	No, between-study comparison
Does statistical analysis suggest that chance is an unlikely explanation for the subgroup difference?	Yes, $P = 0.02$	No, $P = 0.18$	Yes, interaction, $P = 0.003$ in univariable analysis of age 50–69, 70–79, and >80 yr
Did the hypothesis precede rather than follow the analysis, and include a hypothesized direction that was subsequently confirmed?	Yes	One of two studies that enrolled both groups specified the a priori hypothesis	Yes
Was the subgroup hypothesis one of a small number tested?	Yes, one of four	The study that specified a priori tested large number of hypotheses	No, one of 12
Is the subgroup difference consistent across studies and across important outcomes?	Yes, consistent across studies, less so across outcomes	No	Yes, consistent across studies, untested across outcomes
Does external evidence (biological or sociological rationale) support the hypothesized subgroup difference?	Yes, more recruitable lung in which high PEEP should work better in sicker patients	No compelling external evidence supporting subgroup hypothesis	Yes (older persons may have more dietary deficiencies, less exposure to sunlight, thus more vitamin D deficiency)

be rated down for inconsistency is another judgment call; we would argue for not rating down for inconsistency.

1.9. Deciding whether to use estimates from a subgroup analysis: What to do when you are not sure?

Example 3: The systematic review of calcium and vitamin D for fracture prevention included 17 trials in over 50,000 patients. The review authors pooled across all types of fracture (vertebral and nonvertebral) and included studies that randomized patients to intervention groups of calcium or calcium and vitamin D or to a control group receiving neither drug.

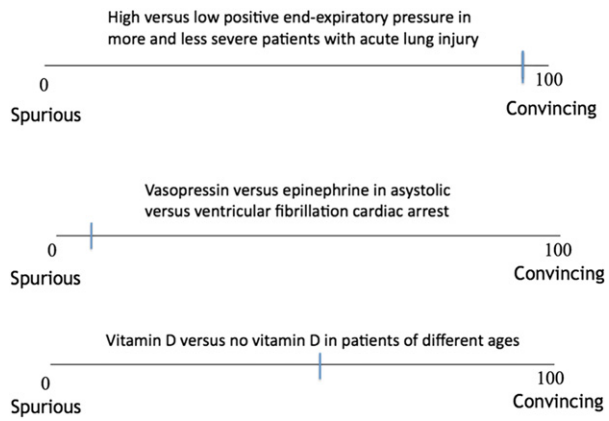

Fig. 6. Credibility of subgroup analyses from three systematic reviews.

The point estimate of the RR was less than 1.0 in all 17 trials; the CI, however, crossed the boundary of no effect in all but three (Fig. 8). The I^2 was 20%, representing little inconsistency in the results of individual studies. The authors nevertheless explored hypotheses (which they specified a priori) about the possibility of there being important inconsistencies between subgroups. In the process, they found an appreciable gradient in effect according to patients' mean age (RRs of 0.97, 0.89, and 0.76 in studies of patients younger than 69, 70–79, and older than 80 years) (Fig. 9).

Applying the seven criteria (Table 1) to this situation, we note that the hypothesis is based on characteristics at randomization, satisfies statistical criteria, was an a priori hypothesis, is consistent with indirect evidence, and is consistent across studies. The hypothesis, however, is supported only by between-study differences, and was one of a dozen a priori hypotheses.

We are therefore left with a subgroup hypothesis of moderate credibility (Fig. 6). A guideline panel is therefore

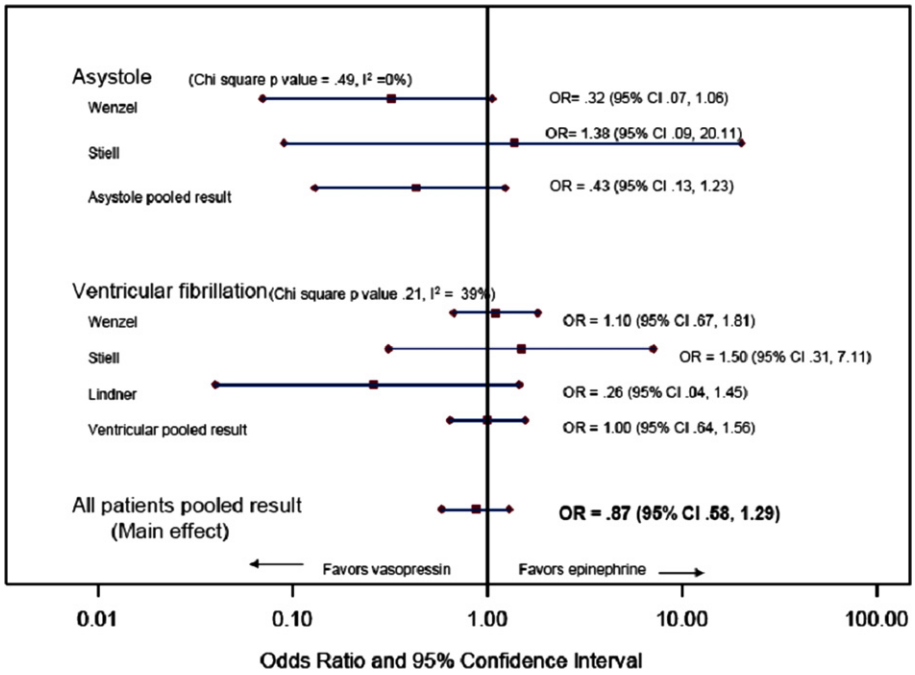

Fig. 7. Vasopressin vs. epinephrine in cardiac arrest.

left with a difficult choice: to offer a recommendation for all persons, or varying recommendations (or varying strength of recommendations) for older and younger people. We would consider either option reasonable.

Fig. 10 highlights issues in the interpretation of subgroup analysis and inconsistency of results when there is an apparent inconsistency among studies. Fig. 10A presents a situation in which there is a little variability in results between studies and no suggestion of a subgroup effect. Systematic review authors and guideline developers will, under these circumstances, present a single pooled estimate and not rate down quality for inconsistency.

In Fig. 10B, authors are persuaded that the subgroup effect is sufficiently credible that it warrants presenting separate evidence summaries for each subgroup. Guideline panels are therefore likely to provide separate recommendations for each subgroup. For neither subgroup will it be necessary to rate down quality for inconsistency.

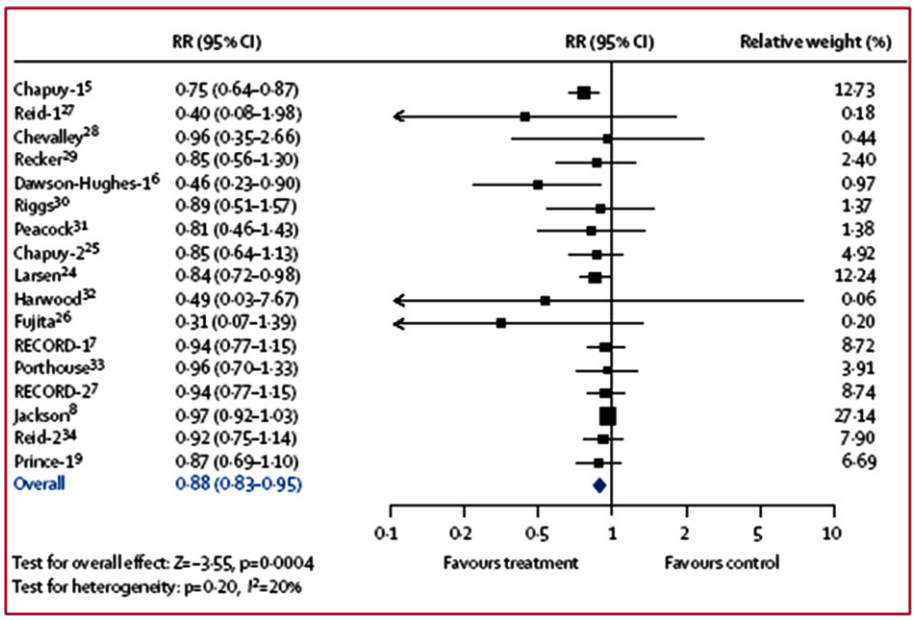

Fig. 8. Fracture reduction with calcium in patients older than 50 yr.

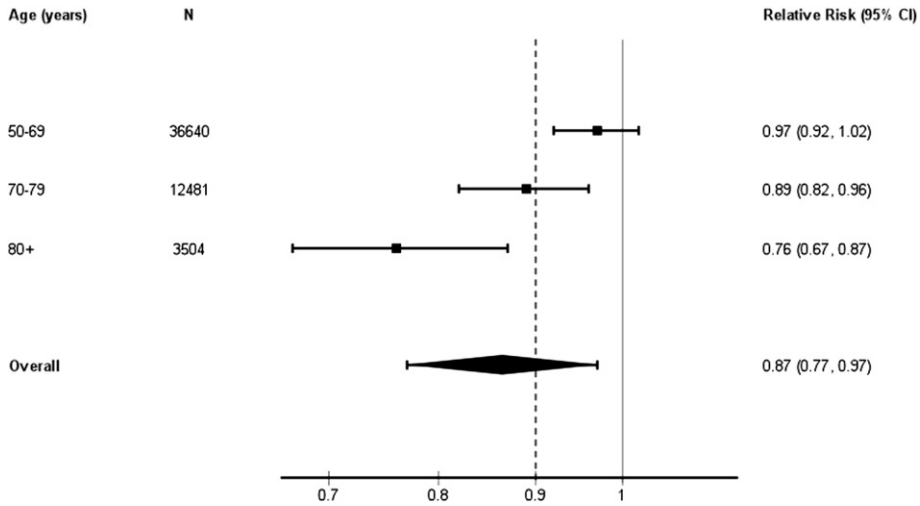

Fig. 9. Apparent fracture reduction with calcium in patients in three age ranges.

Fig. 10C and D depicts situations in which systematic review and guideline authors decide the evidence for a subgroup effect is equivocal. In Fig. 10C, the authors lean toward rejecting the subgroup hypothesis. In this case, they will present a single pooled estimate. Because, however, they are left with appreciable uncertainty as to whom this pooled estimate applies, they may rate down for inconsistency.

In Fig. 10D, systematic review and guideline authors conclude that the subgroup effect is sufficiently credible to warrant presenting separate estimates, but their confidence in this judgment is limited. They present separate effects for each subgroup, but systematic review authors rate down for inconsistency (and guideline panelists may do so as well) because the variability in effects across the subgroups when the subgroup hypothesis may be spurious make them less confident in the estimates of effect they are presenting.

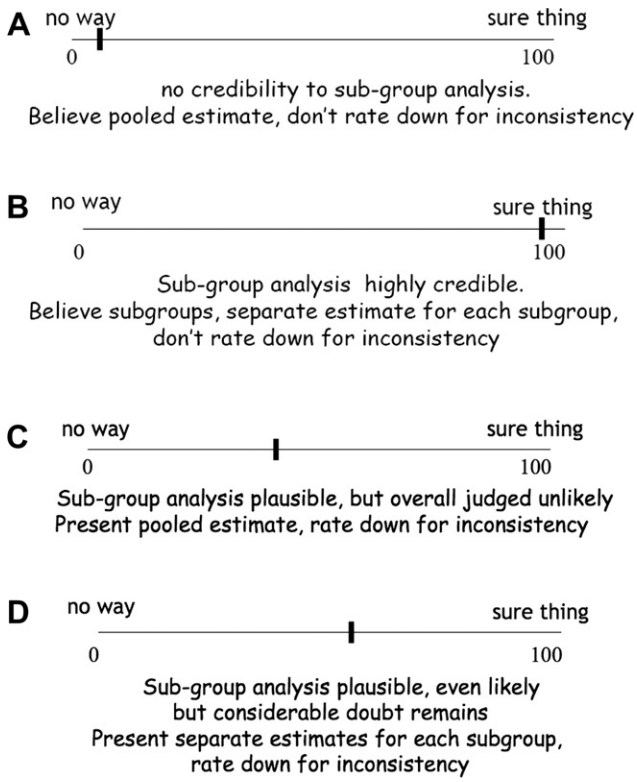

Fig. 10. Interpretation of subgroup analyses of varying credibility.

There is a fifth possibility that the vitamin D example illustrates well. Let us assume that the pooled estimate of effect, and the estimate of effect in one but not all subgroups cross your threshold for recommending a treatment. For instance, assume that a 10% RR reduction was sufficiently large to recommend calcium and vitamin D. Pooled estimates for those aged 70–79 years, those older than 80 years, and the pooled estimate for all studies—but not for those younger than 70 years—are over the chosen threshold (Fig. 9). Now, assume further there are reasons to be skeptical about the subgroup analysis (Fig. 10C and D).

One could argue that the optimal way to deal with this situation would be to present the estimates for all three subgroups, and rate down for inconsistency only for the third (the younger persons). The logic is as follows: for the two older groups of patients, the pooled estimate is above the threshold, and whether one chooses to believe these estimates, or the overall estimate, drug administration is warranted (Fig. 9). Only for the youngest group there is uncertainty: choosing the overall estimate would lead to a recommendation in favor of treatment, choosing the estimate from the subgroup one would recommend against (Fig. 9).

1.10. Conclusion for example 3

What is the appropriate conclusion for the example we have presented? Systematic review and guideline authors might focus on the fact that all point estimates are on the

benefit side, CIs are widely overlapping, the test for heterogeneity is nonsignificant, and the I^2 is low, 20%. Thus, they might conclude that they should ignore the apparent subgroup effect, rating down for inconsistency is unnecessary, and—for the guideline panel—a single recommendation is appropriate for all the age groups (Fig. 10A).

Alternatively, authors may conclude that although they reject the hypothesis that the effect differs in older and younger individuals, doubt remains: perhaps they should provide separate estimates across the three age groups. This would suggest the advisability of rating down for inconsistency: one is uncertain to whom the results apply (Fig. 10C). Uncertainty about to whom the results apply seems particularly troubling in this case: the investigators reported apparent differences in effect between those in long-care institutions and those who are not, and those with lower and higher calcium intake. A full exposition of the issues in this complex consideration would require careful assessment of these other possible subgroup differences, for instance by multivariable meta-regression.

A final possible conclusion is that it is probably best to provide separate estimates for each subgroup effect; nevertheless, uncertainty remains (Fig. 10D). In this case, systematic review and guideline authors may present results separately for the three subgroups (and guideline panels make separate recommendations), and rate down the quality for each recommendation because of inconsistency.

Alternatively, they may use the logic of the fifth situation we have described previously, and rate down the quality for the younger patients only, on the grounds that in the older patients the effect is either as great or greater than for the group as a whole (and the results suggest a statistically significant—and potentially important—effect for the group as a whole) (Fig. 9).

If, as is not the case here, the results suggested important benefit for all the subgroups (but more benefit for one than the others) the situation is analogous to the scenario in Fig. 2 and the flavonoids in hemorrhoid situation we have already discussed. If the benefit is sufficiently large, one might choose not to rate down for inconsistency, the logic being that one is confident of an important effect in all the subgroups, even if one is not confident of its magnitude.

One final consideration: let us assume that one has decided that the subgroup hypothesis is sufficiently credible to present two evidence summaries, one for each subgroup. The subgroup effect has explained some of the variability in results, but it will certainly not explain all the variability. The degree of inconsistency remaining in the results within each subgroup will remain an issue requiring consideration.

References

[1] Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the "number needed to treat"? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol 2002;31(1):72−6.

[2] Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med 2002;21:1575−600.

[3] Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med 1998;17:1923−42.

[4] Guyatt G, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. Grade guidelines: 2. Framing the question. J Clin Epidemiol 2011;64:395−400.

[5] Guyatt G, Oxman A, Vist G, Santesso N, Kunz R, et al. Grade guidelines: 12. Preparing summary of findings tables. J Clin Epidemiol [in press].

[6] Balshem H, Helfand M, Schünemann HJ, Oxman AD, Kunz R, Brozek J, et al. Grade guidelines: 3 Rating the quality of evidence—introduction. J Clin Epidemiol 2011;64:401−6.

[7] Guyatt G, Jaeschke R, Prasad K, Cook D. Summarizing the evidence. In: Guyatt G, Rennie D, Meade M, Cook D, editors. The users' guides to the medical literature: a manual for evidence-based clinical practice. New York, NY: McGraw-Hill; 2008.

[8] Deeks J, Higgins J, Altman D. Analyzing data and undertaking meta-analyses. In: Higgins J, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.0.0. Chichester: Wiley; 2008.

[9] Rucker G, Schwarzer G, Carpenter JR, Schumacher M. Undue reliance on I(2) in assessing heterogeneity may mislead. BMC Med Res Methodol 2008;8:79.

[10] Juni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet 2004;364(9450):2021−9.

[11] Guyatt G, Devereaux PJ, Lexchin J, Stone SB, Yalnizyan A, Himmelstein D, et al. A systematic review of studies comparing health outcomes in Canada and the United States. Open Med 2007;1(1):e27−36.

[12] 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;93:909−20.

[13] Tang BM, et al. Use of calcium or calcium in combination with vitamin D supplementation to prevent fractures and bone loss in people aged 50 years and older: a meta-analysis. Lancet 2007;370(9588):657−66.

[14] Guyatt G, Wyer P, Ioannidis J. When to believe a subgroup analysis. In: Guyatt G, et al, editors. The users' guides to the medical literature: a manual for evidence-based clinical practice. New York, NY: McGraw-Hill; 2008.

[15] Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. JAMA 1994;272(17):1367−71.

[16] Sun X, Briel M, Walter SD, Guyatt GH. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. BMJ 2010;340:c117.

[17] Briel M, Meade M, Mercat A, Brower RG, Talmor D, Walter SD, et al. Higher vs lower positive end-expiratory pressure in patients with acute lung injury and acute respiratory distress syndrome: systematic review and meta-analysis. JAMA 2010;303(9):865−73.

[18] Wyer PC, Perera P, Jin Z, Zhou Q, Cook DJ, Walter SD, et al. Vasopressin or epinephrine for out-of-hospital cardiac arrest. Ann Emerg Med 2006;48(1):86−97.

[19] Wenzel V, Krismer AC, Arntz HR, Sitter H, Stadlbauer KH, Lindner KH, et al. A comparison of vasopressin and epinephrine for out-of-hospital cardiopulmonary resuscitation. N Engl J Med 2004;350(2):105−13.