EXHIBIT 46

Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study

Helene Moustgaard,[1-4] Gemma L Clayton,[5] Hayley E Jones,[5] Isabelle Boutron,[6] Lars Jørgensen,[4] David R T Laursen,[1-4] Mette F Olsen,[4] Asger Paludan-Müller,[4] Philippe Ravaud,[6] Jelena Savović,[5,7], Jonathan A C Sterne,[5,7,8] Julian P T Higgins,[5,7] Asbjørn Hróbjartsson[1-3]

For numbered affiliations see end of the article.

Correspondence to:
H Moustgaard
helene.moustgaard@gmail.com
(or @HeleneMoustgaa1 on Twitter
ORCID 0000-0002-7057-5251)

Additional material is published online only. To view please visit the journal online.

Cite this as: *BMJ* 2020;368:l6802
http://dx.doi.org/10.1136/bmj.l6802

Accepted: 19 November 2019

ABSTRACT

OBJECTIVES

To study the impact of blinding on estimated treatment effects, and their variation between trials; differentiating between blinding of patients, healthcare providers, and observers; detection bias and performance bias; and types of outcome (the MetaBLIND study).

DESIGN

Meta-epidemiological study.

DATA SOURCE

Cochrane Database of Systematic Reviews (2013-14).

ELIGIBILITY CRITERIA FOR SELECTING STUDIES

Meta-analyses with both blinded and non-blinded trials on any topic.

REVIEW METHODS

Blinding status was retrieved from trial publications and authors, and results retrieved automatically from the Cochrane Database of Systematic Reviews. Bayesian hierarchical models estimated the average ratio of odds ratios (ROR), and estimated the increases in heterogeneity between trials, for non-blinded trials (or of unclear status) versus blinded trials. Secondary analyses adjusted for adequacy of concealment of allocation, attrition, and trial size, and explored the association between outcome subjectivity (high, moderate, low) and average bias. An ROR lower than 1 indicated exaggerated effect estimates in trials without blinding.

RESULTS

The study included 142 meta-analyses (1153 trials). The ROR for lack of blinding of patients was 0.91 (95% credible interval 0.61 to 1.34) in 18 meta-analyses with patient reported outcomes, and 0.98 (0.69 to 1.39) in 14 meta-analyses with outcomes reported by blinded observers. The ROR for lack of blinding of healthcare providers was 1.01 (0.84 to 1.19) in 29 meta-analyses with healthcare provider decision outcomes (eg, readmissions), and 0.97 (0.64 to 1.45) in 13 meta-analyses with outcomes reported by blinded patients or observers. The ROR for lack of blinding of observers was 1.01 (0.86 to 1.18) in 46 meta-analyses with subjective observer reported outcomes, with no clear impact of degree of subjectivity. Information was insufficient to determine whether lack of blinding was associated with increased heterogeneity between trials. The ROR for trials not reported as double blind versus those that were double blind was 1.02 (0.90 to 1.13) in 74 meta-analyses.

CONCLUSION

No evidence was found for an average difference in estimated treatment effect between trials with and without blinded patients, healthcare providers, or outcome assessors. These results could reflect that blinding is less important than often believed or meta-epidemiological study limitations, such as residual confounding or imprecision. At this stage, replication of this study is suggested and blinding should remain a methodological safeguard in trials.

WHAT IS ALREADY KNOWN ON THIS TOPIC

Blinding is an established methodological procedure in randomised clinical trials

Empirical estimates of the expected degree of bias in trials due to lack of blinding can help interpret trial results (eg, in a systematic review or clinical guideline) and plan future trials

Previous meta-epidemiological studies have reported variable estimates of the effect of blinding, with little discussion of who was blinded and the type of outcome

WHAT THIS STUDY ADDS

This large meta-epidemiological study of 142 Cochrane meta-analyses found no evidence that lack of blinding of patients, healthcare providers, or outcome assessors had an impact on effect estimates in randomised clinical trials, on average

This finding does not support the importance of blinding and is inconsistent with some previous studies; but it is consistent with several other smaller meta-epidemiological studies

The results indicate that blinding, on average, could be less important than previously believed, or could reflect limitations in the meta-epidemiological approach, such as confounding and misclassification; replication of the study is recommended and, at present, no change to methodological practice is suggested

Introduction

A randomised clinical trial is the most reliable method for assessing the effect of therapeutic interventions.[1] Results of clinical trials underpin evidence based clinical practice and decisions made by regulatory agencies, either directly or as part of a meta-analysis. However, results of randomised clinical trials might be biased[2]—for example, by systematic differences between the care provided to participants or systematic differences in the behaviour of participants, in the intervention and comparison groups (performance bias); or by systematic differences between these groups in the way in which outcomes are assessed (detection bias). Blinding (sometimes called masking) of patients, healthcare providers, and outcome assessors is intended to prevent such bias.

Blinding is used in some form in about 60% of trials.[3] However, blinding of patients and healthcare

providers is sometimes not possible owing to the type of interventions being tested (eg, psychotherapy). In other instances, blinding might not be applied owing to logistical challenges. Historically, use of placebo control interventions and blinding procedures was closely linked to early development of the randomised trial. Blinding has been an established methodological principle since around 1950.[4]

Various meta-epidemiological studies have investigated the effect of blinding on estimated intervention effects.[5 6] Such studies collate large numbers of meta-analyses of randomised trials, compare the results of blinded and non-blinded trials within meta-analyses, and then combine estimated within-meta-analysis differences across meta-analyses.[6] Estimates of the average impact of blinding have shown considerable variation between studies.[7] These studies mostly dealt with several types of bias simultaneously, and their analyses had conceptual and methodological limitations. Comparison of double blind trials with trials that are not double blinded is problematic, because the double blind concept is ambiguous.[8 9] This ambiguity is especially clear in non-pharmacological trials, and the comparison does not enable separation of performance bias and detection bias. To date, all meta-epidemiological studies of blinding have relied exclusively on information provided by trial publications, where inadequate reporting of blinding is common. Only one study took into account by whom outcomes were reported.[10]

A more comprehensive analysis of the impact of blinding in randomised trials is important. Designers of trials have to consider whether spending resources on blinding is worthwhile. Users of trial information (eg, consumers, researchers conducting systematic reviews, and guideline developers) must assess the risk of bias due to incomplete blinding.

We conducted a meta-epidemiological study to estimate the separate effects of blinding patients, healthcare providers, and outcome assessors on the results of randomised clinical trials. We also estimated the impact of different types of blinding on between-study heterogeneity.

Methods

Identification of meta-analyses for inclusion

We sought meta-analyses that included at least one trial with blinding of patients, healthcare providers, or outcome assessors (that is, observers) and at least one trial without blinding of the same groups. We refer to these as informative meta-analyses. To identify these, we screened all 1042 Cochrane reviews published or updated between 1 February 2013 and 18 February 2014 (Cochrane Database of Systematic Reviews, issue 2, 2013). We used Cochrane risk of bias tool[2] assessments to select potentially informative meta-analyses suitable for further data extraction. Specifically, we examined the first listed meta-analysis in the review's table of contents with an observer reported outcome and a difference between trials in the risk of bias score for detection bias (high *v* low or high *v* unclear risk); and with a patient reported or healthcare provider decision outcome (outcomes determined by clinical decisions—eg, readmissions or need for surgical intervention) and a difference between trials in the risk of bias score for performance bias.

The screening process identified 395 potentially informative meta-analyses. Of these, 226 provided information on blinding of outcome assessors and 169 on blinding of patients or healthcare providers. For pragmatic reasons, we selected for further study a random subsample of 120 meta-analyses from the former set, but retained all of the latter set, giving a total of 289 potentially informative meta-analyses (full details are in the appendix).

Data retrieval and extraction

Trial publications (and any corresponding protocols/methods publications) were retrieved for each trial in each potentially informative meta-analysis. When publications could not readily be retrieved, we requested a copy from Cochrane review authors. For trials published after 1999 and where the blinding status of trial participants was unclear we contacted authors by email, asking for information on the blinding status of all groups within the trial.

We read the full text of publications in languages known to us (English, Danish, French, German, and Spanish). For publications in other languages (eg, Chinese) we based data extraction on any English language abstract, but did not attempt translation of the full text.

Data on basic trial characteristics and information on blinding status were extracted manually from trial publications. Trial results were extracted automatically from the Cochrane Database of Systematic Reviews through the Archie database interface: number of patients in intervention and control groups, for binary outcomes the number of events, and for measurement scale outcomes the means and standard deviations. We also automated extraction of the name of the Cochrane review group, and review authors' risk of bias assessments for the domains "allocation concealment" and "incomplete outcome data."

Assessment of blinding status

We assessed the blinding status of patients, healthcare providers, and outcome assessors using a modified algorithm derived from that of Akl and colleagues[11] (full details are given in the appendix). The algorithm entailed contacting trial authors (for trials published after 1999) when there was insufficient information on blinding in the trial publications. We defined blinding as a lack of awareness by patients, healthcare providers, or outcome assessor of the intervention status of individual patients throughout the trial.

We coded healthcare providers as blinded if all staff groups involved in patient treatment and care were described as "blinded" (eg, doctors and nurses, or all staff), and as non-blinded if all, or a subgroup, were described as "non-blinded" (eg, surgeons). Staff responsible for healthcare provider decision outcomes

doi: 10.1136/bmj.l6802 | *BMJ* 2020;368:l6802 | the bmj

were thus also covered by the blinding status of healthcare providers.

We differentiated between definitive information on blinding status (definitely yes/definitely no) based on explicit description or contact with trial authors, and assessments based on other information in publications (probably yes/probably no). For instance, for drug trials using a placebo control and described as "double blind" or "triple blind," patients, healthcare providers, and outcome assessors were all classified as blinded (probably yes), unless stated explicitly otherwise. For trials with no mention of "placebo," "double dummy," "double blinding," "triple blinding," "single blinding," or similar, all trial groups were classified as non-blinded (probably no), unless stated explicitly otherwise. Assessment of blinding status was made by two observers independently (AP-M, DRTL, LJ, MFO, HM, or AH), and any differences were resolved by discussion between the two. When we did not receive a reply from authors, or where we did not attempt contact, the blinding status was recorded as unclear.

When making a final determination of whether meta-analyses were informative, and for the purposes of our analyses, we compared trials that had relevant parties recorded as having "definitely no," "probably no," or "unclear" blinding with those that had relevant parties coded as "definitely yes" or "probably yes." After detailed assessment of blinding status, 189 of the 289 meta-analyses were classified as informative.

Classifications and exclusions

Classification of interventions as experimental and control was based on descriptions in the trial publications, except when the review clearly labelled the comparator as "placebo," "control," "standard care," or "treatment as usual," in which case we followed the labelling used by the review authors and classified these interventions as controls. To ensure consistent comparisons of estimated bias across meta-analyses, we excluded those meta-analyses in which intervention classifications were unclear.

Outcome measures were classified as observer reported, patient reported (via interviewer or directly recorded by patients), healthcare provider decision outcomes, or mixed (in instances where the outcome was a mixture of more than one category—eg, both patient and observer reported elements). We excluded meta-analyses of trials that did not all have the same type of outcome (eg, patient reported) unless there was an informative subset of trials with the same type of outcome.

Observer reported outcomes were subdivided into four outcomes: objective—all cause mortality, objective—other than total mortality (eg, automatised non-repeatable laboratory tests), subjective—pure observation (eg, assessment of radiographs), and subjective—interactive (eg, assessment of clinical status). Subjective observer reported outcomes were scored 1-3 according to the degree of subjectivity (that is, the extent to which determination of the outcome depended on the judgment of the observer, with 1 indicating a low degree of subjectivity). The scoring of subjectivity was done by two observers (HM and MFO) independently and masked to any results of trials or meta-analyses, with any differences resolved by discussion. Box 1 shows examples of outcomes and subjectivity scores.

Meta-analyses were classified according to whether the outcome was measured in the trials based on an underlying hypothesis of benefit (eg, degree of pain measured based on the hypothesis that the intervention lowers pain) or of harm (eg, frequency of allergic reactions measured based on the hypothesis that the intervention could cause an increase). Classification of outcomes according to clinical area and type of experimental and comparison interventions was conducted to facilitate comparisons with an earlier meta-epidemiological study.[12] We further categorised experimental interventions as alternative/complementary or conventional medicine, to facilitate comparison with a systematic review of trials randomising patients to blinded and unblinded substudies.[13]

We excluded trials with binary outcomes, in which no or all participants had the outcome event, and trials with continuous outcomes, where the required information for calculating the standardised mean difference was missing. We also excluded trials included in more than one meta-analysis with the same outcome, if the meta-analyses were to be included in the same meta-epidemiological analysis. Such trials were removed at random until the trial occurred only within one meta-analysis. After removal of individual trials, some meta-analyses were no longer informative. The final study database contained 142 meta-analyses with a total of 1153 trials.

Data analysis

All main analyses were prespecified. In our main analyses, which included only meta-analyses with outcomes measured based on a hypothesis of benefit, we differentiated between types of bias (detection bias and performance bias) and category of person blinded (patient, healthcare provider, and outcome assessor). We performed five main analyses, quantifying the average association between estimates of treatment effect and lack of blinding:

Box 1: Examples of subjectivity scoring of trial outcomes

- Subjectivity score 1 (low degree of subjectivity): heart rate, forced expiratory volume in first second (FEV_1), cotinine saliva dipstick assay
- Subjectivity score 2 (medium degree of subjectivity): superficial surgical site infection, recurrence of varicose veins, tooth prosthesis failure
- Subjectivity score 3 (high degree of subjectivity): change in global measure of cognition, Barthel index score (of ability to perform activities of daily living), Hamilton depression scale score

- (Ia) Blinding of patients in trials with patient reported outcomes (considering a combination of detection bias and performance bias)
- (Ib) Blinding of patients in trials with blinded observer reported outcomes (considering performance bias)
- (IIa) Blinding of healthcare providers in trials with healthcare provider decision outcomes (considering a combination of detection bias and performance bias)
- (IIb) Blinding of healthcare providers in trials with blinded observers or patients assessing the outcome (considering performance bias)
- (III) Blinding of outcome assessors (that is, observers) in trials with subjective outcomes (considering detection bias).

We did not primarily focus on trials with objective outcomes, such as all cause mortality, because we did not suspect any marked effect of blinding in such trials. We conducted univariable analyses for each contrast in blinding status using all informative meta-analyses for that characteristic.

Intervention effects for binary outcomes were modelled as log odds ratios and coded such that an odds ratio of less than 1 indicated a beneficial intervention effect. For continuous outcomes, the standardised mean difference and corresponding standard error were used and coded such that a standardised mean difference of less than zero meant a beneficial intervention effect.

We quantified differences in intervention effects, comparing non-blinded trials with blinded trials of each type using ratios of odds ratios: $ROR=OR_{non\text{-}blinded}/OR_{blinded}$. Bayesian hierarchical models for meta-epidemiological research, developed by Welton and colleagues, were used to estimate the average bias associated with lack of each type of blinding (ROR), the average variability in this bias within a meta-analysis (quantified by κ, the standard deviation increase in heterogeneity between trials), and variability in average bias between meta-analyses (quantified by ϕ, the standard deviation in mean bias between meta-analyses).[14]

The model thus enabled us to explore the average degree of bias, and also whether the bias differs (eg, in direction) between meta-analyses (that is, the importance of blinding might depend on the clinical scenario) and between trials (that is, the importance of blinding might depend on factors related to the singular trial, even within similar clinical scenarios).

The analyses were carried out using Markov chain Monte Carlo simulations in WinBUGS version 1.4.3. Vague prior distributions were assumed for all parameters (see appendix for more details). We modelled continuous and binary data simultaneously, assuming a mixture of normal and binomial likelihoods but modelling the underlying bias on the same scale. This method required re-expressing standardised mean differences as odds ratios.[15] To reduce risk of spurious findings, we defined a lower threshold of at least 10 meta-analyses for conducting an analysis.

To study the impact of subjectivity scores on the average difference in intervention effect associated with blinding outcome assessors, we extended the model of Welton et al[14] to incorporate a three level categorical covariate (low *v* moderate *v* high degree of subjectivity) at the meta-analysis level.

In sensitivity analyses, we excluded trials with a classification of blinding status as "unclear" from the analyses. Secondary analyses were stratified by outcome type (eg, objective outcomes and subtypes).

Confounding by other flaws in trial design was assessed in multivariable analyses by re-running each of the five main analyses with adjustment in the model for concealment of the allocation sequence, incomplete outcome data (attrition), trial size, and blinding status of patients. The blinding status of patients was only included in the analysis of outcome assessor blinding (III). We adjusted for each of these characteristics in separate analyses. We did not include combinations of the covariates.

We also conducted post hoc subgroup analyses according to type of outcome data (continuous *v* binary) and type of comparator (active control *v* inactive control), calculated the impact of concealment of the allocation sequence on estimated treatment effects, and repeated the main analyses using an alternative label-invariant meta-epidemiological model, proposed recently by Rhodes et al.[16] This model removes the constraint that intervention effects are at least as variable among the non-blinded trials as among the blinded trials within each meta-analysis, but was not available when we wrote our protocol.

Finally, to facilitate comparison of our results with previous meta-epidemiological studies we also compared trials described by trial authors as "double blind" or "triple blind" with those not described in this way.

Patient and public involvement

Patients and members of the public were not involved in the research because it was designed to answer a methodological challenge that was not directly dependent on patient priorities, experiences, or participant preferences. The methodological expertise required to plan the study, analyse the results, and write the manuscript was dependent on specialist knowledge and we did not try to identify patients or members of the public with this training to work with.

Results

The final study database contained 142 meta-analyses with a total of 1153 trials. Figure 1 shows the flow of data through the study, from screening to final dataset. We contacted the trial author for 54 (5%) of the 1153 trials in the dataset. In 28 instances the authors replied (response rate 52%), and the fraction of trials with unclear blinding status was thereby reduced from 95/1153 (8%) to 67/1153 (6%).

doi: 10.1136/bmj.l6802 | BMJ 2020;368:l6802 | the bmj

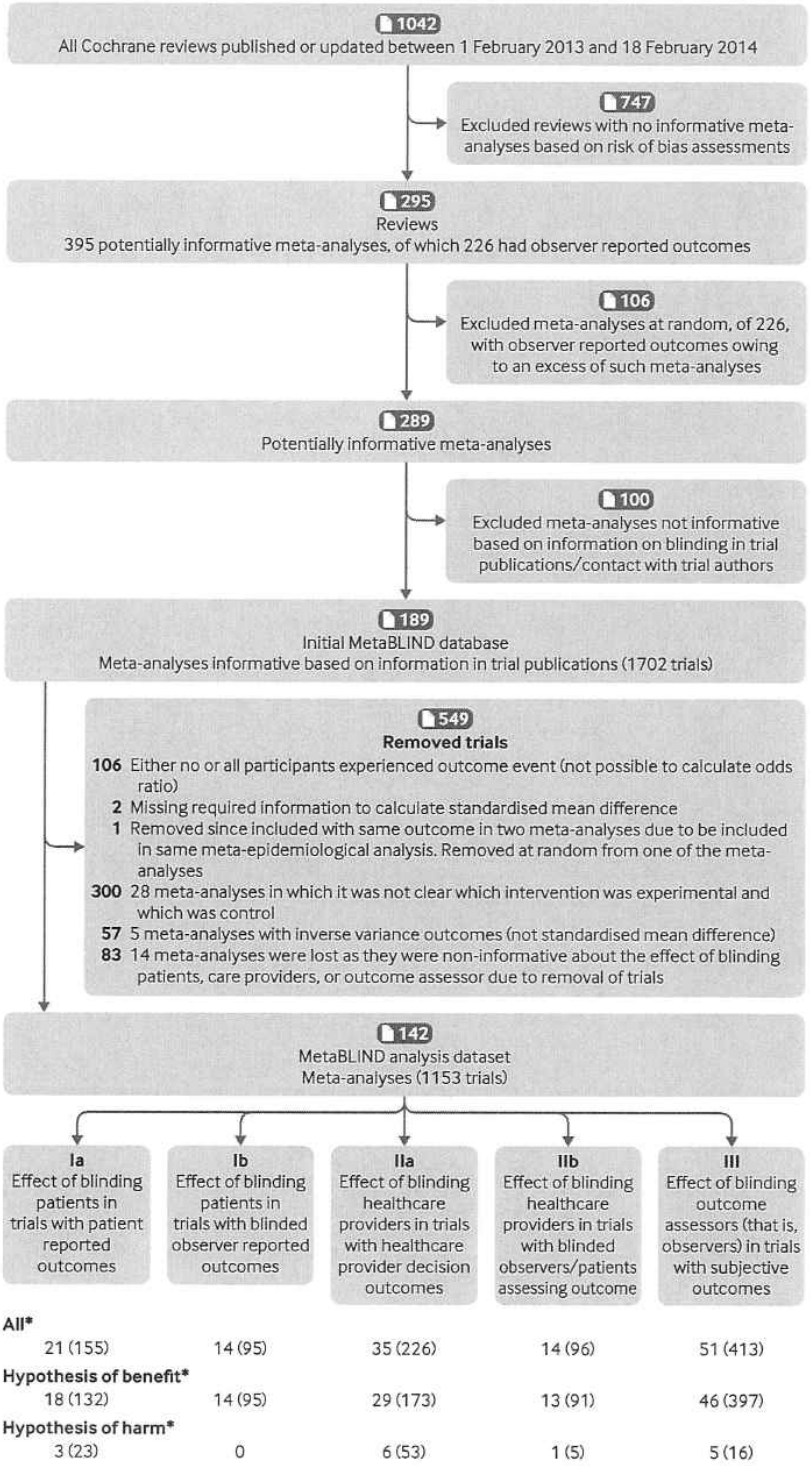


Fig 1 | Study flow diagram. *Meta-analyses contributing with trials that had outcome measures categorised as "mixed" (that is, it was not possible to classify them as patient reported, healthcare provider decision, or observer reported because they contained elements from more than one of these types) were not counted. Mixed outcome trials did not contribute to the main analyses

Table 1 | Characteristics of meta-analyses and trials included for the overall dataset and main analyses

Characteristics	Overall dataset		Main analyses*									
			Ia		Ib		IIa		IIb		III	
	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials
Outcome measures according to clinical area	142	1153	18	132	14	95	29	173	13	91	46	397
Adverse events of treatment	22 (15.5)	129 (11.2)	0	0	0	0	0	0	0	0	0	0
All cause mortality	7 (4.9)	143 (12.4)	0	0	2 (14.3)	27 (28.4)	0	0	2 (15.4)	27 (29.7)	0	0
Cause specific mortality	1 (0.7)	11 (1.0)	0	0	0	0	0	0	0	0	1 (2.2)	11 (2.8)
Clinician assessed outcomes (eg, body mass index, blood pressure, lung function, infant weight)	12 (8.5)	95 (8.2)	0	0	1 (7.1)	11 (11.6)	1 (3.4)	3 (1.7)	0	0	11 (23.9)	92 (23.2)
Composite endpoint (including mortality or major morbidity)	2 (1.4)	16 (1.4)	0	0	2 (14.3)	12 (12.6)	0	0	1 (7.7)	7 (7.7)	1 (2.2)	9 (2.3)
Global improvement	3 (2.1)	14 (1.2)	0	0	2 (14.3)	5 (5.3)	0	0	2 (15.4)	5 (5.5)	2 (4.3)	12 (3.0)
Laboratory reported outcomes (eg, blood components, tissue analysis, urinalysis)	5 (3.5)	45 (3.9)	0	0	1 (7.1)	2 (2.1)	0	0	0	0	1 (2.2)	4 (1.0)
Lifestyle outcomes (including diet, exercise, smoking)	5 (3.5)	100 (8.7)	1 (5.6)	2 (1.5)	0	0	0	0	0	0	3 (6.5)	63 (15.9)
Major morbidity event (including myocardial infarction, stroke, haemorrhage)	5 (3.5)	44 (3.8)	0	0	3 (21.4)	24 (25.3)	0	0	3 (23.1)	24 (26.4)	5 (10.9)	44 (11.1)
Mental health outcomes (including cognitive function, depression and anxiety scores)	7 (4.9)	61 (5.3)	2 (11.1)	9 (6.8)	1 (7.1)	4 (4.2)	0	0	0	0	5 (10.9)	52 (13.1)
Other outcomes (not classified elsewhere)	15 (10.6)	145 (12.6)	5 (27.8)	79 (59.8)	1 (7.1)	2 (2.1)	4 (13.8)	16 (9.2)	2 (15.4)	4 (4.4)	5 (10.9)	48 (12.1)
Pain (extent of pain a patient is experiencing)	5 (3.5)	17 (1.5)	3 (16.7)	8 (6.1)	0	0	0	0	1 (7.7)	7 (7.7)	1 (2.2)	2 (0.5)
Perinatal outcomes	5 (3.5)	34 (2.9)	0	0	0	0	1 (3.4)	2 (1.2)	0	0	1 (2.2)	9 (2.3)
Pregnancy outcomes	8 (5.6)	28 (2.4)	0	0	0	0	1 (3.4)	3 (1.7)	0	0	6 (13.0)	23 (5.8)
Quality of life (including ability to perform physical, daily, and social activities)	3 (2.1)	19 (1.6)	2 (11.1)	6 (4.5)	0	0	0	0	0	0	1 (2.2)	13 (3.3)
Radiological outcomes (including radiograph abnormalities, ultrasound, magnetic resonance imaging results)	2 (1.4)	11 (1.0)	0	0	1 (7.1)	8 (8.4)	0	0	1 (7.7)	8 (8.8)	2 (4.3)	11 (2.8)
Resource use (including cost, hospital stay duration, number of procedures)	19 (13.4)	133 (11.5)	0	0	0	0	19 (65.5)	133 (76.9)	0	0	0	0
Surgical and device related outcomes	4 (2.8)	20 (1.7)	0	0	0	0	3 (10.3)	16 (9.2)	0	0	1 (2.2)	4 (1.0)
Symptoms or signs of illness or condition	6 (4.2)	35 (3.0)	5 (27.8)	28 (21.2)	0	0	0	0	1 (7.7)	9 (9.9)	0	0
Withdrawals/dropouts/compliance	6 (4.2)	53 (4.6)	0	0	0	0	0	0	0	0	0	0
Type of experimental intervention												
Pharmacological	95 (66.9)	728 (63.1)	12 (66.7)	48 (36.4)	10 (71.4)	74 (77.9)	19 (65.5)	121 (69.9)	10 (76.9)	78 (85.7)	25 (54.3)	195 (49.1)
Surgical	3 (2.1)	12 (1.0)	1 (5.6)	4 (3.0)	0	0	0	0	0	0	1 (2.2)	4 (1.0)
Psychosocial, behavioural, or educational	17 (12.0)	204 (17.7)	1 (5.6)	42 (31.8)	3 (21.4)	17 (17.9)	3 (10.3)	10 (5.8)	1 (7.7)	2 (2.2)	9 (19.6)	101 (25.4)
Other	27 (19.0)	209 (18.1)	4 (22.2)	38 (28.8)	1 (7.1)	4 (4.2)	7 (24.1)	42 (24.3)	2 (15.4)	11 (12.1)	11 (23.9)	97 (24.4)
Field of experimental intervention												
Conventional medicine	137 (96.5)	1100 (95.4)	17 (94.4)	127 (96.2)	14 (100.0)	95 (100.0)	29 (100.0)	173 (100.0)	12 (92.3)	84 (92.3)	44 (95.7)	368 (92.7)
Alternative/complementary medicine	5 (3.5)	53 (4.6)	1 (5.6)	5 (3.8)	0	0	0	0	1 (7.7)	7 (7.7)	2 (4.3)	29 (7.3)
Type of comparison intervention												
Placebo or no treatment	57 (40.1)	442 (38.3)	8 (44.4)	36 (27.3)	1 (7.1)	11 (11.6)	12 (41.4)	47 (27.2)	2 (15.4)	16 (17.6)	17 (37.0)	160 (40.3)
Other inactive (standard care)	38 (26.8)	452 (39.2)	4 (22.2)	76 (57.6)	7 (50.0)	55 (57.9)	9 (31.0)	84 (48.6)	5 (38.5)	46 (50.5)	17 (37.0)	176 (44.3)
Active comparison	47 (33.1)	259 (22.5)	6 (33.3)	20 (15.2)	6 (42.9)	29 (30.5)	8 (27.6)	42 (24.3)	6 (46.2)	29 (31.9)	12 (26.1)	61 (15.4)
Hypothesis of benefit	114 (80.3)	971 (84.2)	18 (100.0)	132 (100.0)	14 (100.0)	95 (100.0)	29 (100.0)	173 (100.0)	13 (100.0)	91 (100.0)	46 (100.0)	397 (100.0)

doi: 10.1136/bmj.l6802 | BMJ 2020;368:l6802 | the bmj

Table 1 | Continued

Characteristics	Overall dataset		Main analyses*									
			Ia		Ib		IIa		IIb		III	
	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials	Meta-analyses	Trials
Observer reported outcome†	68 (47.9)	640 (55.5)	0	0	14 (100.0)	95 (100.0)	0	0	10 (76.9)	73 (80.2)	46 (100.0)	397 (100.0)
All cause mortality	11 (16.2)	170 (26.6)	0	0	2 (14.3)	27 (28.4)	0	0	2 (20.0)	27 (37.0)	0	0
Other objective	4 (5.9)	39 (6.1)	0	0	1 (7.1)	2 (2.1)	0	0	0	0	0	0
Subjective	53 (77.9)	431 (67.3)	0	0	11 (78.6)	66 (69.5)	0	0	8 (80.0)	46 (63.0)	46 (100.0)	397 (100.0)
Binary or measurement scale outcome												
Binary	110 (77.5)	885 (76.8)	9 (50.0)	42 (31.8)	11 (78.6)	78 (82.1)	25 (86.2)	151 (87.3)	11 (84.6)	82 (90.1)	32 (69.6)	289 (72.8)
Continuous	31 (21.8)	265 (23.0)	8 (44.4)	87 (65.9)	3 (21.4)	17 (17.9)	4 (13.8)	22 (12.7)	2 (15.4)	9 (9.9)	14 (30.4)	108 (27.2)
Inverse variance	1 (0.7)	3 (0.3)	1 (5.6)	3 (2.3)	0	0	0	0	0	0	0	0
Year of publication of trial (median, IQR)	—	2003 (1996-2008)	2008 (2004-11)	2005 (1997-2009)	2011 (2005-12)	2005 (1998-2009)	2009 (2005-11)	2000 (1996-2008)	2010 (2003-12)	2002 (1995-09)	2010 (2003-12)	2003 (1996-2008)
Sample size of meta-analysis/trial (median, IQR)	768 (293-2025)	106 (50-270)	706 (163-1314)	133 (78-234)	809 (173-2402)	102 (43-300)	838 (323-3103)	139 (63-483)	1085 (421-3621)	149 (61-370)	599 (289-1361)	82 (40-207)

IQR=interquartile range. Results are shown as number (%) unless stated otherwise.
*Ia=Effect of blinding patients in trials with patient reported outcomes; Ib=effect of blinding patients in trials with blinded observer reported outcomes; IIa=effect of blinding healthcare providers in trials with healthcare provider decision outcomes; IIb=effect of blinding healthcare providers in trials with blinded observers or patients assessing the outcome; III=effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes.
†Judged for the whole meta-analysis, not outcomes in individual trials.

Appendix table 1 shows the proportions of trials classified as definitely yes and probably yes.

Table 1 shows characteristics of the 142 meta-analyses and 1153 trials included in the dataset. The median year of trial publication was 2003 (interquartile range 1996-2008), and the median sample size was 768 (293-2025) patients for meta-analyses and 106 (50-270) for trials. Of the 1153 trials included in the analysis dataset, 1112 (96%) had a parallel trial design and 753 (65%) were drug trials. Full details are given in appendix table 1.

Various methodological characteristics were strongly associated across trials. For instance, trials in which the outcome assessor was blinded were more likely to have adequate allocation concealment (odds ratio 3.0, 95% confidence interval 2.2 to 4.0) and complete outcome data (2.0, 1.5 to 2.8). Trials reporting that patients were blinded were more likely to report that the outcome assessor was blinded (75.0, 38.6 to 145.8). Full details are shown in appendix tables 2 and 3. Figure 2 presents results for each of the five main analyses (Ia, Ib, IIa, IIb, III). Forest plots of the meta-analyses are shown in appendix figure 1.

For the effect of blinding patients in trials with patient reported outcomes (analysis Ia), 18 informative meta-analyses with a hypothesis of benefit contained 132 trials. Patient blinding was assessed as probably yes or definitely yes in 33 trials (25%). The average ROR was 0.91 (95% credible interval 0.61 to 1.34). The average standard deviation increase in heterogeneity between trials among non-blinded trials was very imprecisely estimated and is presented in figure 2 and appendix table 4, together with implied 95% predictive intervals for the ROR in a single trial, to facilitate interpretation. For the effect of blinding patients in trials with blinded observer reported outcomes (analysis Ib), 14 informative meta-analyses with a hypothesis of benefit contained 95 trials. Patient blinding was assessed as probably yes or definitely yes in 57 (60%) of these. The average ROR was 0.98 (95% credible interval 0.69 to 1.39).

For the effect of blinding healthcare providers in trials with healthcare provider decision outcomes (analysis IIa), 29 informative meta-analyses with a hypothesis of benefit contained 173 trials. Healthcare provider blinding was assessed as probably yes or definitely yes in 93 of these trials (54%). The average ROR was 1.01 (95% credible interval 0.84 to 1.19). For the effect of blinding healthcare providers in trials with blinded observers or patients assessing the outcome (analysis IIb), 13 informative meta-analyses with a hypothesis of benefit contained 91 trials. Healthcare provider blinding was assessed as probably yes or definitely yes in 61 trials (67%). The average ROR was 0.97 (95% credible interval 0.64 to 1.45).

For the effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes (analysis III), 46 informative meta-analyses with a hypothesis of benefit contained 397 trials. Outcome assessor blinding was assessed as probably or definitely yes in 199 of these trials (50%). The average ROR was 1.01

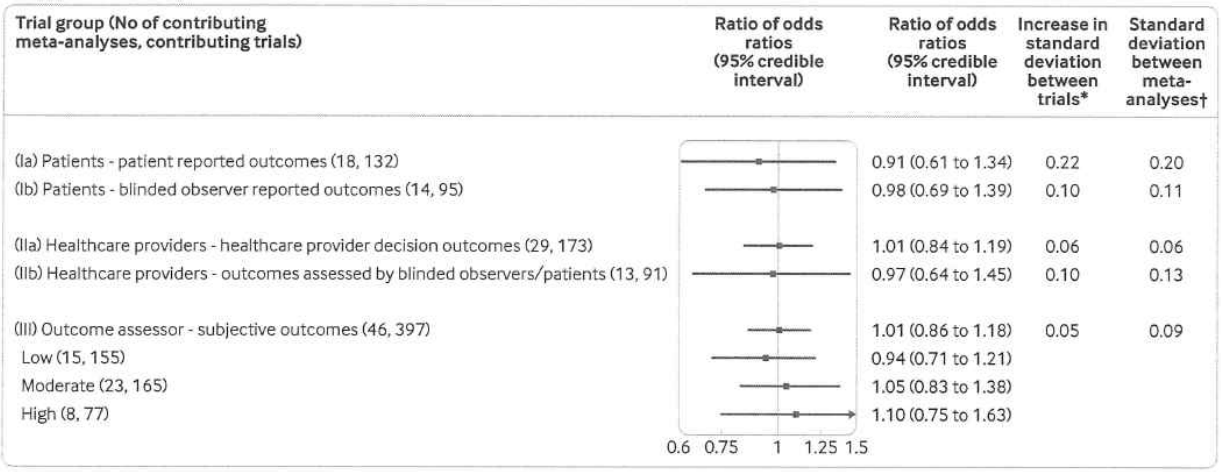


Fig 2 | Estimated ratios of odds ratios and effects on heterogeneity associated with blinding status of patients, healthcare providers, and outcome assessors. Unadjusted analyses. *Increase in standard deviation between trials: (Ia) 0.22 (95% credible interval 0.02 to 0.60), (Ib) 0.10 (0.01 to 0.30), (IIa) 0.06 (0.01 to 0.30), (IIb) 0.10 (0.01 to 0.59), (III) 0.05 (0.01 to 0.22). †Standard deviation between meta-analyses: (Ia) 0.20 (95% credible interval 0.01 to 0.74), (Ib) 0.11 (0.01 to 0.55), (IIa) 0.06 (0.01 to 0.26), (IIb) 0.13 (0.01 to 0.82), (III) 0.09 (0.01 to 0.31)

(95% credible interval 0.86 to 1.18). In the additional analysis in which we explored the impact of the level of subjectivity of the outcome, we estimated average RORs of 0.94 (0.71 to 1.21), 1.05 (0.83 to 1.38), and 1.10 (0.75 to 1.63) for outcomes with low, moderate, and high degree of subjectivity, respectively.

For each of the five main analyses, separate adjustment for concealment of the allocation sequence, attrition, and trial size did not materially change the result (table 2). Estimated increases in heterogeneity between trials and estimates of variability between meta-analyses in average bias also did not change substantially, compared with the unadjusted main analyses.

Analyses comparing trials described as "double blind" (or "triple blind") with those not so described, or with an unclear status, did not show any effect when they included meta-analyses with any type of outcome (ROR 0.99, 95% credible interval 0.86 to 1.09), nor when they included only meta-analyses with subjective observer reported outcomes and a hypothesis of benefit (1.11, 0.86 to 1.44; table 3). Exclusion of trials with an unclear blinding status from the unadjusted main analyses did not change the results substantially (table 3).

Results of secondary analyses looking separately at the effect of blinding patients, healthcare providers, or outcome assessors across different types of outcomes are shown in appendix table 5. For example, an analysis based on observer reported outcomes classified as objective also showed little evidence of an effect of outcome assessor blinding status (ROR 0.94, 95% credible interval 0.61 to 1.26; meta-analyses with a hypothesis of benefit only).

A pre-planned repetition of the main analyses based only on trials scored as definitely yes versus trials scored as definitely no proved unfeasible due to insufficient numbers of meta-analyses (appendix table 5). A post hoc analysis indicated about 10% exaggeration of the odds ratio in trials without adequate concealment of the allocation sequence (table 3). We report the results of other post hoc analyses for type of outcome (continuous *v* binary) and type of comparator (active control *v* inactive control) in table 3.

Results for the five main analyses repeated using the alternative, label-invariant, model of Rhodes et al[16] are presented in appendix table 6. The estimates of ROR and of heterogeneity between meta-analyses in bias from both models were similar. Results for heterogeneity between trials were not directly comparable to those for the main model, but indicated a possible increase in heterogeneity among blinded trials, although again the parameter estimates were very imprecise.

Discussion

We found no evidence of a difference, on average, in estimated treatment effects between randomised clinical trials with and without blinding of patients, between trials with and without blinding of healthcare providers, and between trials with and without blinding of outcome assessors. In all instances the credible intervals were wide, including both considerable difference and no difference. The same pattern was found when comparing trials that were double blind with those that were not. Our findings of an increase in heterogeneity between trials are inconclusive, owing to a lack of information.

Strengths and challenges of the study

The main strengths and originality of our study were that blinding was analysed according to the type of person blinded and due consideration given to the type of outcome. Analysis in this way allowed a separation of the two main types of blinding related bias

doi: 10.1136/bmj.l6802 | *BMJ* 2020;368:l6802 | the bmj

Table 2 | Adjusted analyses. Data are outcome measure (95% credible interval) unless stated otherwise

No of meta-analyses, trials	Adjusted for patient blinding			Adjusted for allocation concealment			Adjusted for incomplete outcome data			Adjusted for trial size		
	ROR	Φ	κ	ROR	Φ	κ	ROR	Φ	κ	ROR	Φ	κ
(Ia) Effect of blinding patients in trials with patient reported outcomes												
18, 132	NA			0.91 (0.61 to 1.35)	0.20 (0.02 to 0.74)	0.21 (0.01 to 0.61)	0.91 (0.63 to 1.31)	0.17 (0.01 to 0.70)	0.18 (0.01 to 0.60)	0.89 (0.59 to 1.29)*	0.18 (0.02 to 0.74)	0.18 (0.01 to 0.60)
(Ib) Effect of blinding patients in trials with blinded observer reported outcomes												
14, 95	NA			1.07 (0.74 to 1.56)	0.11 (0.01 to 0.57)	0.10 (0.01 to 0.57)	1.08 (0.72 to 1.58)	0.10 (0.01 to 0.52)	0.13 (0.01 to 0.72)	0.99 (0.69 to 1.39)	0.10 (0.01 to 0.54)	0.10 (0.01 to 0.57)
(IIa) Effect of blinding healthcare providers in trials with healthcare provider decision outcomes												
29, 173	NA			1.03 (0.84 to 1.23)	0.07 (0.01 to 0.29)	0.06 (0.01 to 0.28)	0.98 (0.80 to 1.17)	0.06 (0.01 to 0.28)	0.07 (0.01 to 0.30)	1 (0.83 to 1.19)	0.06 (0.01 to 0.27)	0.06 (0.01 to 0.29)
(IIb) Effect of blinding healthcare providers in trials with blinded observers/patients assessing the outcome												
13, 91	NA			1.03 (0.67 to 1.54)	0.13 (0.01 to 0.80)	0.10 (0.01 to 0.60)	1.07 (0.69 to 1.64)	0.12 (0.01 to 0.77)	0.09 (0.01 to 0.60)	0.98 (0.63 to 1.44)	0.13 (0.01 to 0.82)	0.09 (0.01 to 0.58)
(III) Effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes												
46, 397	1.03 (0.87 to 1.23)	0.10 (0.01 to 0.32)	0.06 (0.01 to 0.22)	1.04 (0.89 to 1.23)	0.10 (0.01 to 0.36)	0.05 (0.01 to 0.21)	1.02 (0.87 to 1.19)	0.08 (0.01 to 0.33)	0.05 (0.01 to 0.19)	1.03 (0.88 to 1.21)	0.10 (0.01 to 0.34)	0.06 (0.01 to 0.25)

NA=not applicable; ROR=ratio of odds ratios; κ=standard deviation increase in heterogeneity between trials; Φ=standard deviation in mean bias between meta-analyses.
*One meta-analysis (three trials) was removed, which did not specify the size of the trial owing to the format given in the review.

(performance and detection bias) and enabled a comprehensive analysis that was less reliant on the way in which authors used the phrase "double blind." Also, we had a low proportion of trials with unclear blinding status, partly because we attempted to contact the trial authors. We restricted the main analyses to outcomes measured, based on a hypothesis of benefit, and ensured that interventions considered experimental in our analyses were also regarded as experimental in the individual trials.

The specificity of the comparisons limited the number of trials and meta-analyses that could be included in individual analyses, which restricted the precision of estimated differences between trials with and without the various types of blinding. We planned our sample size pragmatically, primarily based on results of comparisons within trials.[13 17-19] Formal power calculations were published after we had planned our study.[20]

Meta-epidemiological studies are observational and so estimated effects of trial characteristics could be confounded. We adjusted for predefined variables such as allocation concealment, attrition, trial size, and blinding status of patients. Concurrent adjustment for a combination of factors was not feasible, and confounding by unknown or unmeasured factors could have affected results.

Confounding by other methodological characteristics can be expected to exaggerate the estimated effect of lack of blinding, rather than cancel it. Nevertheless, attenuation of the estimated effect of blinding by confounding cannot be ruled out. For instance, more pragmatically conducted trials within a meta-analysis (those with the broadest inclusion criteria and with least control of treatment adherence) could be less likely to have used blinding and could have resulted in less beneficial treatment effects than more explanatory trials. The consequence would be to move the estimated ROR towards 1.

Blinding could have less impact in trials comparing an experimental intervention with an active comparator (that is, not compared with placebo, no treatment, or standard care). Type of comparator, however, did not seem to affect the analysis of outcome assessor blinding, and too few informative meta-analyses precluded additional analyses. Possibly, blinding could have less impact in trials that aim to determine an intention-to-treat effect than in trials aiming to determine a per protocol effect. We did not explore whether the impact of blinding differed according to inferential goal or type of analysis.

Blinding could be lost during the course of a trial,[21] which would tend to attenuate the apparent differences between blinded and non-blinded trials. Other factors to consider are a possibly larger impact of non-reporting bias on blinded trials, and misclassification (despite our intensive efforts to classify correctly the blinding status of patients, healthcare providers, and outcome assessors). In general, non-differential misclassification would bias our results towards no impact of lack of blinding.

Table 3 | Secondary analyses. Data are outcome measure (95% credible interval) unless stated otherwise

	No of meta-analyses, trials	ROR	Φ	κ
Lack of double blinding or unclear double blinding (v double blind)				
All outcomes	94, 722	0.99 (0.86 to1.09)	0.07 (0.01 to 0.29)	0.06 (0.01 to 0.18)
Benefit	74, 583	1.02 (0.90 to 1.13)	0.06 (0.01 to 0.27)	0.07 (0.01 to 0.19)
Harms	20, 139	0.64 (0.38 to 1.04)	0.15 (0.01 to 0.89)	0.13 (0.01 to 1.23)
Observer reported outcomes: benefit	36, 374	1.04 (0.84 to 1.25)	0.14 (0.01 to 0.57)	0.08 (0.01 to 0.23)
Subjectively assessed observer reported outcomes: benefit	27, 221	1.11 (0.86 to 1.44)	0.13 (0.01 to 0.61)	0.09 (0.01 to 0.42)
Patient reported outcomes: benefit	13, 53	0.89 (0.57 to 1.40)	0.15 (0.01 to 0.83)	0.12 (0.01 to 0.88)
Healthcare provider decision outcomes: benefit	24, 147	0.98 (0.79 to 1.19)	0.07 (0.01 to 0.31)	0.07 (0.01 to 0.36)
Repeat of the main analyses excluding trials with unclear blinding status				
(Ia) Effect of blinding patients in trials with patient reported outcomes	16, 116	1.10 (0.72 to 1.69)	0.19 (0.02 to 0.76)	0.23 (0.02 to 0.61)
(Ib) Effect of blinding patients in trials with blinded observer reported outcomes	14, 94	1.00 (0.70 to 1.44)	0.11 (0.01 to 0.58)	0.10 (0.01 to 0.60)
(IIa) Effect of blinding healthcare providers in trials with healthcare provider decision outcomes	28, 160	0.97 (0.77 to 1.18)	0.08 (0.01 to 0.36)	0.07 (0.01 to 0.39)
(IIb) Effect of blinding healthcare providers in trials with blinded observers/patients assessing the outcome	13, 90	0.96 (0.64 to 1.45)	0.14 (0.01 to 0.82)	0.10 (0.01 to 0.68)
(III) Effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes	43, 365	1.01 (0.85 to 1.20)	0.11 (0.01 to 0.35)	0.06 (0.01 to 0.25)
Main analysis by type of outcome				
(III) Effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes: continuous*	14, 108	dSMD 0.02 (−0.22 to 0.26)	0.07 (0.01 to 0.37)	0.07 (0.01 to 0.31)
(III) Effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes: binary*	32, 289	1.01 (0.85 to 1.20)	0.11 (0.01 to 0.37)	0.06 (0.01 to 0.23)
Main analyses by type of control intervention				
(III) Effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes: active control*	12, 61	1.01 (0.64 to 1.55)	0.12 (0.01 to 0.70)	0.10 (0.01 to 0.56)
(III) Effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes: inactive control (placebo/no treatment/standard care)*	34, 336	1.01 (0.85 to 1.21)	0.10 (0.01 to 0.36)	0.06 (0.01 to 0.23)
Inadequate (or unclear) concealment of the allocation sequence (v adequate)				
(Ia) Effect of blinding patients in trials with patient reported outcomes†	13, 116	0.95 (0.68 to 1.29)	0.11 (0.01 to 0.63)	0.10 (0.01 to 0.51)
(IIa) Effect of blinding healthcare providers in trials with healthcare provider decision outcomes†	22, 154	0.90 (0.72 to 1.12)	0.09 (0.01 to 0.35)	0.07 (0.01 to 0.32)
(III) Effect of blinding outcome assessors (that is, observers) in trials with subjective outcomes†	40, 349	0.88 (0.76 to 1.02)	0.07 (0.01 to 0.27)	0.08 (0.01 to 0.30)

ROR=ratio of odds ratios; κ=standard deviation increase in heterogeneity between trials; Φ=standard deviation in mean bias between meta-analyses; dSMD=difference in standardised mean difference.
*The prespecified minimum of 10 meta-analyses for analysis to be feasible was met only in analysis III.
†Analyses include meta-analyses from each of the datasets used in the main analyses that were informative for the impact of inadequate (or unclear) concealment of the allocation sequence. The numbers of informative meta-analyses in analyses Ib and IIb did not meet the prespecified minimum of 10 meta-analyses for analysis to be feasible.

The generalisability of our results could be affected by the sampling strategy inherent in a meta-epidemiological approach. Thus, inclusion of only meta-analyses containing both blinded and non-blinded trials excludes situations where all trials are blinded (as blinding is considered of paramount importance) or, conversely, areas where all trials tend to be non-blinded. Similarly, review authors might be more likely to include both blinded and non-blinded trials in a meta-analysis when there is no clear difference in effect estimates between the two.

Our estimation of average bias (ROR) was robust with regard to choice of statistical model.[14 16] The same applied to our analyses of heterogeneity in bias between meta-analyses. The model restriction embedded in the additive model by Welton and colleagues,[14] used for our main analyses, however, implies that between-trial heterogeneity among non-blinded trials can only increase (or remain unchanged). We reanalysed our data with an alternative model not restricted by this assumption,[16] which was not available when we planned our study. The reanalysis indicated a possible decrease in heterogeneity among non-blinded trials, although estimates were imprecise, and results were also consistent with a considerable increase in heterogeneity between trials. We interpret this result cautiously, to imply that there was insufficient information to determine whether lack of blinding was associated with increased heterogeneity between trials. Few direct comparisons have been published

doi: 10.1136/bmj.l6802 | BMJ 2020;368:l6802 | the bmj

between the newly developed label-invariant model[16] and the additive model[14] used in our study and in most large meta-epidemiological studies.[12 22] Analyses of the ROBES study database based on the additive model indicated an increase in heterogeneity between trials among trials with inadequate or unclear concealment of allocation, whereas the label-invariant model indicated a decrease.[16]

Other studies

Systematic reviews of meta-epidemiological studies[7 23] identified four studies (comparisons within meta-analyses) estimating the impact of blinding patients, three studies estimating the impact of blinding trial personnel, and four studies estimating the impact of blinding outcome assessors. In all instances, blinding had surprisingly little effect.[7 23] Two additional recent studies partly confirmed this pattern: an analysis of physiotherapy trials[24] found little evidence of an impact of blinding of patients or of outcome assessors, and a study of oral health trials[25] found no evidence of an impact of blinding of outcome assessors, though some evidence of a moderate effect of patient blinding.

By contrast, three systematic reviews of within-trial comparisons for 51 trials with both blinded and non-blinded outcome assessment found that blinding had a clear effect.[17-19] For example, non-blinded outcome assessors of subjective[26] outcomes exaggerated odds ratios by 36%, on average.[17] Similarly, a systematic review of 12 trials randomising patients to blinded and non-blinded substudies reported a pronounced bias due to lack of patient blinding in complementary/alternative medicine trials with patient reported outcomes, exaggerating effect sizes by 0.56 standard deviations.[13] Such comparisons within trials have no major risk of confounding. The trial design is rare, however, so to what extent the results could be generalised is not clear.

Results of meta-epidemiological studies comparing double blind trials with trials without (or unclear) double blinding have shown noticeable variation.[7] A systematic review by Page and colleagues found an overall 8% exaggeration of odds ratios in trials without double blinding (although confidence intervals overlapped no effect),[7] and an exaggeration of 23% when outcomes were subjective.[7 12]

Mechanisms and implications

Clarification of the circumstances in which blinding is important in trials, and an empirical assessment of direction and degree of bias, have important and direct implications for the design of future trials, for interpretation of trial results, and for instructions on how to assess risk of bias when conducting systematic reviews. Clarification is also pertinent to the current debate on the balance between reliability and relevance of unblinded patient reported outcome measures (PROMS),[27 28] and the relative importance of blinded explanatory trials versus unblinded pragmatic trials.[29]

Convincing theoretical reasons lead us to expect both detection and performance bias in non-blinded trials. Experimental psychology backs the notion that expectations and interest tend to shape human evaluations.[30 31] Comparisons within trials[13 17-19] provide strong evidence that in specific settings lack of blinding in trials causes considerable bias. Exactly what characterises these settings is unclear, however. We suggest that replication of our study would be valuable, as would updates of the systematic reviews of comparisons within trials, and exploration of the conditions under which blinding is more, or less, important.

Meta-epidemiological studies are often used to assess empirically dimensions of bias in randomised trials, but they could themselves be biased. For example, meta-epidemiological studies of allocation concealment have disclosed an unexpected dependence of impact on type of outcome.[12] Theoretically, impact of allocation concealment should not depend on the subjectivity of outcomes.[7 32] We suggest careful consideration of the risk of confounding and of bias, such as bias due to misclassification of methodological characteristics or due to erroneous identification of treatments as experimental and control, in meta-epidemiological studies.[33]

Blinding has been considered an essential methodological precaution in trials for decades. We did not expect to find that our study does not firmly underpin standard methodological practice. Further, our results are coherent with other meta-epidemiological studies that have reported similar results. The implication seems to be that either blinding is less important (on average) than often believed, that the meta-epidemiological approach is less reliable, or that our findings can, to some extent, be explained by lack of precision. At present, we suggest that assessors of the risk of bias in trials included in a systematic review continue to deal with the implications of lack of blinding for risk of bias, as is done in version 2 of the Cochrane risk of bias tool.[34]

In conclusion, we found no evidence of a difference, on average, in estimated treatment effect between randomised clinical trials with blinded and non-blinded patients, between trials with blinded and non-blinded healthcare providers, and between trials with blinded and non-blinded outcome assessors. The apparent lack of a major average effect of blinding on estimated treatment effects is surprising to us and is at odds with methodological standard practices. We are unclear to what extent our results show that blinding is less important than previously believed, show the limitations of the meta-epidemiological approach (eg, residual confounding), or show a lack of precision in the comparisons made. Until our study has been replicated, and we have a clearer understanding of which types of trials are susceptible to bias associated with lack of blinding, we suggest that blinding remains an important methodological safeguard in trials in which it is feasible.

AUTHOR AFFILIATIONS
[1]Centre for Evidence-Based Medicine Odense (CEBMO), Odense University Hospital, Kløvervænget 10, DK-5000 Odense C, Denmark

[2]Open Patient data Explorative Network (OPEN), Odense University Hospital, Odense, Denmark

[3]Department of Clinical Research, University of Southern Denmark, Odense, Denmark

[4]Nordic Cochrane Centre, Copenhagen, Denmark

[5]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[6]Cochrane France, Hôpital Hôtel-Dieu, Paris, France

[7]The National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care West (NIHR CLAHRC West) at University Hospitals Bristol NHS Foundation Trust, Bristol, UK

[8]NIHR Bristol Biomedical Research Centre, University of Bristol, Bristol, UK

We thank former Cochrane editor in chief David Tovey for providing us with access to the Cochrane Database of Systematic Reviews, and the Nordic Cochrane Centre, particularly Rasmus Moustgaard, for enabling automatic data extraction from the database.

Contributors: AH and HM conceived and organised the study, interpreted the results, and drafted the manuscript. HM also extracted data. GLC analysed the data, interpreted results, and drafted the manuscript. HEJ, JS, IB, PR, JPTH, and JACS conceived the study, interpreted the results, and drafted the manuscript. LJ, DRTL, AP-M, and MFO extracted data and drafted the manuscript. HM is guarantor. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Funding: This study received no specific funding. GLC was funded by a PhD studentship from the Medical Research Council (MRC) Hubs for Trials Methodology Research. HEJ was supported by an MRC Career Development Award in Biostatistics (MR/M014533/1). JS and JPTH are supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care West (CLAHRC West). JACS and JPTH are NIHR senior investigators (NF-SI-0611-10168 and NF-SI-0617-10145, respectively), are supported by NIHR Bristol Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol, and are members of the MRC Integrative Epidemiology Unit at the University of Bristol. The views expressed are those of the author(s) and not necessarily those of the National Health Service, the NIHR, or the UK Department of Health and Social Care.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: Not required.

Data sharing: Dataset available from the corresponding author after a post-publication period of 1 year allowing time for follow-up projects.

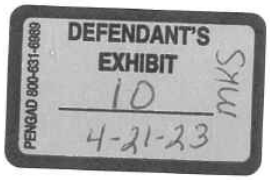
The lead author affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Dissemination to participants and related patient and public communities: We plan to present our findings at national and international scientific meetings. We also plan to use social media outlets to disseminate findings. We will consider the implication of our findings for assessing the risk of bias in results of randomised trials using version 2 of the Cochrane risk of bias assessment tool.

1 Berlin JA, Golub RM. Meta-analysis as evidence: building a better pyramid. *JAMA* 2014;312:603-5. doi:10.1001/jama.2014.8167
2 Higgins JPT, Altman DG, Gøtzsche PC, et al, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ* 2011;343:d5928. doi:10.1136/bmj.d5928
3 Chan A-W, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. *Lancet* 2005;365:1159-62. doi:10.1016/S0140-6736(05)71879-1
4 Kaptchuk TJ. Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bull Hist Med* 1998;72:389-433. doi:10.1353/bhm.1998.0159
5 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. doi:10.1001/jama.1995.03520290060030
6 Sterne JAC, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. *Stat Med* 2002;21:1513-24. doi:10.1002/sim.1184
7 Page MJ, Higgins JPT, Clayton G, Sterne JAC, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. *PLoS One* 2016;11:e0159267. doi:10.1371/journal.pone.0159267
8 Haahr MT, Hróbjartsson A. Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors. *Clin Trials* 2006;3:360-5. doi:10.1177/1740774506069153
9 Devereaux PJ, Manns BJ, Ghali WA, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. *JAMA* 2001;285:2000-3. doi:10.1001/jama.285.15.2000
10 Nüesch E, Reichenbach S, Trelle S, et al. The importance of allocation concealment and patient blinding in osteoarthritis trials: a meta-epidemiologic study. *Arthritis Rheum* 2009;61:1633-41. doi:10.1002/art.24894
11 Akl EA, Sun X, Busse JW, et al. Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. *J Clin Epidemiol* 2012;65:262-7. doi:10.1016/j.jclinepi.2011.04.015
12 Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med* 2012;157:429-38. doi:10.7326/0003-4819-157-6-201209180-00537
13 Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. *Int J Epidemiol* 2014;43:1272-83. doi:10.1093/ije/dyu115
14 Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. *J R Stat Soc Ser A Stat Soc* 2009;172:119-36. doi:10.1111/j.1467-985X.2008.00548.x
15 Chinn S. A simple method for converting an odds ratio to effect size for use in meta-analysis. *Stat Med* 2000;19:3127-31. doi:10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M
16 Rhodes KM, Mawdsley D, Turner RM, Jones HE, Savović J, Higgins JPT. Label-invariant models for the analysis of meta-epidemiological data. *Stat Med* 2018;37:60-70. doi:10.1002/sim.7491
17 Hróbjartsson A, Thomsen AS, Emanuelsson F, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ* 2012;344:e1119. doi:10.1136/bmj.e1119
18 Hróbjartsson A, Thomsen AS, Emanuelsson F, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. *CMAJ* 2013;185:E201-11. doi:10.1503/cmaj.120744
19 Hróbjartsson A, Thomsen ASS, Emanuelsson F, et al. Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *Int J Epidemiol* 2014;43:937-48. doi:10.1093/ije/dyt270
20 Giraudeau B, Higgins JPT, Tavernier E, Trinquart L. Sample size calculation for meta-epidemiological studies. *Stat Med* 2016;35:239-50. doi:10.1002/sim.6627
21 Bello S, Moustgaard H, Hróbjartsson A. Unreported formal assessment of unblinding occurred in 4 of 10 randomized clinical trials, unreported loss of blinding in 1 of 10 trials. *J Clin Epidemiol* 2017;81:42-50. doi:10.1016/j.jclinepi.2016.08.002
22 Savović J, Turner RM, Mawdsley D, et al. Association between risk-of-bias assessments and results of randomized trials in Cochrane reviews: the ROBES meta-epidemiologic study. *Am J Epidemiol* 2018;187:1113-22. doi:10.1093/aje/kwx344
23 Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. *J Clin Epidemiol* 2016;77:24-37. doi:10.1016/j.jclinepi.2016.04.005
24 Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in physical therapy trials and its association with treatment effects: a meta-epidemiological study. *Am J Phys Med Rehabil* 2017;96:34-44. doi:10.1097/PHM.0000000000000521
25 Saltaji H, Armijo-Olivo S, Cummings GG, Amin M, da Costa BR, Flores-Mir C. Influence of blinding on treatment effect size estimate in randomized controlled trials of oral health interventions. *BMC Med Res Methodol* 2018;18:42. doi:10.1186/s12874-018-0491-0

26 Moustgaard H, Bello S, Miller FG, Hróbjartsson A. Subjective and objective outcomes in randomized clinical trials: definitions differed in methods publications and were often absent from trial reports. *J Clin Epidemiol* 2014;67:1327-34. doi:10.1016/j.jclinepi.2014.06.020
27 Ghimire P, Hasegawa H, Kalyal N, Hurwitz V, Ashkan K. Patient-reported outcome measures in neurosurgery: a review of the current literature. *Neurosurgery* 2018;83:622-30. doi:10.1093/neuros/nyx547
28 Claessen FM, Mellema JJ, Stoop N, Lubberts B, Ring D, Poolman RW. Influence of priming on patient-reported outcome measures: a randomized controlled trial. *Psychosomatics* 2016;57:47-56. doi:10.1016/j.psym.2015.09.005
29 Ware JH, Hamel MB. Pragmatic trials--guides to better patient care?*N Engl J Med* 2011;364:1685-7. doi:10.1056/NEJMp1103502
30 Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev Gen Psychol* 1998;2:175-220. doi:10.1037/1089-2680.2.2.175
31 Rosenthal R. On the social psychology of the psychological experiment: the experimenter's hypothesis as unintended determinant of experimental results. *Am Sci* 1963;51:268-83.
32 Higgins JPT, Ramsay C, Reeves BC, et al. Issues relating to study design and risk of bias when including non-randomized studies in systematic reviews on the effects of interventions. *Res Synth Methods* 2013;4:12-25. doi:10.1002/jrsm.1056
33 Moustgaard H, Jones HE, Savović J, et al. Ten questions to consider when interpreting results of a meta-epidemiological study – the MetaBLIND study as a case. *Res Synth Meth* 2019;1-15. doi:10.1002/jrsm.1392
34 Sterne JAC, Savović J, Page MJ, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ* 2019;366:l4898. doi:10.1136/bmj.l4898

Supplementary information: Appendix



ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1277–1282

Journal of Clinical Epidemiology

GRADE SERIES - SHARON STRAUS, RACHEL CHURCHILL AND SASHA SHEPPERD, GUEST EDITORS

GRADE guidelines: 5. Rating the quality of evidence—publication bias

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

[a]*Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[b]*Department of Medicine, McMaster University, Hamilton, Ontario L8N 3Z5, Canada*
[c]*Norwegian Knowledge Centre for the Health Services, PO Box 7004, St Olavs plass, 0130 Oslo, Norway*
[d]*Knowledge and Encounter Research Unit, Mayo Clinic, Rochester, MN, USA*
[e]*Basel Institute of Clinical Epidemiology, University Hospital Basel, Hebelstrasse 10, 4031 Basel, Switzerland*
[f]*Iberoamerican Cochrane Center-Servicio de Epidemiología Clínica y Salud Pública and CIBER de Epidemiología y Salud Pública (CIBERESP), Hospital de Sant Pau, Universidad Autónoma de Barcelona, Barcelona 08041, Spain*
[g]*Center for Evidence-based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL, 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]*Department of Veterans Affairs, QUERI Program, Office of Research and Development, Washington, DC, USA*
[k]*Department of Medicine, Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*
[l]*Durham VA Center for Health Services Research in Primary Care, Duke University Medical Center, Durham, NC 27705, USA*
[m]*German Cochrane Center, Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, 79104 Freiburg, Germany*
[n]*Division of Pediatric Hematology and Oncology, Department of Pediatric and Adolescent Medicine, University Medical Center Freiburg, 79106 Freiburg, Germany*
[o]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, USA*
[p]*Department of Medicine, State University of New York at Buffalo, New York, NY, USA*

Accepted 5 January 2011; Published online 30 July 2011

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 a body of evidence is associated with a high risk of publication bias. Even when individual studies included in best-evidence summaries have a low risk of bias, publication bias can result in substantial overestimates of effect. Authors should suspect publication bias when available evidence comes from a number of small studies, most of which have been commercially funded. A number of approaches based on examination of the pattern of data are available to help assess publication bias. The most popular of these is the funnel plot; all, however, have substantial limitations. Publication bias is likely frequent, and caution in the face of early results, particularly with small sample size and number of events, is warranted.

Keywords: GRADE; Quality of evidence; Publication bias; Funnel plot; Conflict of interest; Pharmaceutical industry

1. Introduction

In four previous articles in our series describing the GRADE system of rating the quality of evidence and grading the strength of recommendations, we have described the process of framing the question, introduced GRADE's approach to rating the quality of evidence, and dealt with the possibility of rating down quality for study limitations

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, 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.2011.01.011

Key points

- Empirical evidence shows that, in general, studies with statistically significant results are more likely to be published than studies without statistically significant results ("negative studies").
- Systematic reviews performed early, when only few initial studies are available, will overestimate effects when "negative" studies face delayed publication. Early positive studies, particularly if small in size, are suspect.
- Recent revelations suggest that withholding of "negative" results by industry sponsors is common. Authors of systematic reviews should suspect publication bias when studies are uniformly small, particularly when sponsored by the industry.
- Empirical examination of patterns of results (e.g., funnel plots) may suggest publication bias but should be interpreted with caution.

(risk of bias). This fifth article deals with the another of the five categories of reasons for rating down the quality of evidence: publication bias. Our exposition relies to some extent on prior work addressing issues related to publication bias [1]; we did not conduct a systematic review of the literature relating to publication bias.

Even if individual studies are perfectly designed and executed, syntheses of studies may provide biased estimates because systematic review authors or guideline developers fail to identify studies. In theory, the unidentified studies may yield systematically larger or smaller estimates of beneficial effects than those identified. In practice, there is more often a problem with "negative" studies, the omission of which leads to an upward bias in estimate of effect. Failure to identify studies is typically a result of studies remaining unpublished or obscurely published (e.g., as abstracts or theses)—thus, methodologists have labeled the phenomenon "publication bias."

An informative systematic review assessed the extent to which publication of a cohort of clinical trials is influenced by the statistical significance, perceived importance, or direction of their results [2]. It found five studies that investigated these associations in a cohort of registered clinical trials. Trials with positive findings were more likely to be published than trials with negative or null findings (odds ratio: 3.90; 95% confidence interval [CI]: 2.68, 5.68). This corresponds to a risk ratio of 1.78 (95% CI: 1.58, 1.95), assuming that 41% of negative trials are published (the median among the included studies, range = 11–85%). In absolute terms, this means that if 41% of negative trials are published, we would expect that 73% of positive trials would be published. Two studies assessed time to publication and showed that trials with positive findings tended to be published after 4–5 years compared with those with negative findings, which were published after 6–8 years. Three studies found no statistically significant association between sample size and publication. One study found no statistically significant association between either funding mechanism, investigator rank, or sex and publication.

2. Publication bias vs. selective reporting bias

In some classification systems, reporting bias has two subcategories: selective outcome reporting, with which we have dealt in the previous article in the series, and publication bias. However, all the sources of bias that we have considered under study limitations, including selective outcome reporting, can be addressed in single studies. In contrast, when an entire study remains unreported and reporting is related to the size of the effect—publication bias—one can assess the likelihood of publication bias only by looking at a group of studies [2–7]. Currently, we follow the Cochrane approach and consider selective reporting bias as an issue in risk of bias (study limitations). This issue is currently under review by the Cochrane Collaboration, and both Cochrane and GRADE may revise this in future.

3. Variations in publication bias

The results of a systematic review will be biased if the sample of studies included is unrepresentative—whether the studies not included are published or not. Thus, biased conclusions can result from an early review that omits studies with delayed publication—a phenomenon sometimes termed "lag bias" [8]. Either because authors do not submit studies with what they perceive as uninteresting results to prominent journals or because of repeated rejection at such journals, a study may end up published in an obscure journal not indexed in major databases and not identified in a less-than-comprehensive search. Authors from non-English speaking countries may submit their negative studies to local journals not published in English [9,10]; these will inevitably be missed by any review that restricts itself to English-language journals. Negative studies may be published in some form (theses, book chapters, compendia of meeting abstract submissions—sometimes referred to as "gray literature") that tend to be omitted from systematic reviews without comprehensive searching [11].

With each of these variations of publication bias, there is a risk of overestimating the size of an effect. However, the importance of unpublished studies, non-English language publication and gray literature are difficult to predict for individual systematic reviews.

One may have a mirror image phenomenon to the usual publication bias: a study may be published more than once, with different authors and changes in presentation that

make the duplication difficult to identify, and potentially lead to double counting of results within systematic reviews [12–15].

Meta-analyses of *N*-acetylcysteine for preventing contrast-induced nephropathy demonstrate a number of these phenomena [16]. Randomized trials reported only in abstract form in major cardiology journals showed smaller effects than trials fully published. Of those trials published, the earlier published studies showed larger effects than the later published studies. Studies with positive results were published in journals with higher impact factors than studies with negative conclusions. Systematic reviews proved vulnerable to these factors, included published studies more often than abstracts, and conveyed inflated estimates of treatment effect. Table 1 presents a number of ways that selective or nonpublication can bias the results of a best-evidence summary classified according to the phase of the publication process.

4. Bigger dangers of publication bias in reviews with small studies

The risk of publication bias may be higher for reviews that are based on small randomized controlled trials (RCTs) [17–19]. RCTs including large numbers of patients are less likely to remain unpublished or ignored and tend to provide more precise estimates of the treatment effect, whether positive or negative (i.e., showing or not showing a statistically significant difference between intervention and control groups). Discrepancies between results of meta-analyses of small studies and subsequent large trials may occur as often as 20% of the time [20], and publication bias may be a major contributor to the discrepancies [21].

5. Large studies are not immune

Although large studies are more likely to be published, sponsors who are displeased with the results may delay or even suppress publication [14,22,23]. Furthermore, they may publish in journals with limited readership studies that, by their significance, warrant publication in the highest profile medical journals. They may also succeed in obscuring results using strategies that are scientifically unsound. The following example illustrates all these phenomena.

Salmeterol Multicentre Asthma Research Trial (SMART) was a randomized trial that examined the impact of salmeterol or placebo on a composite outcome of respiratory-related deaths and life-threatening experiences. In September 2002, after a data monitoring committee review of 25,858 randomized patients showed a nearly significant increase in the primary outcome in the salmeterol group, the sponsor, GlaxoSmithKline (GSK), terminated the study. Deviating from the original protocol, GSK submitted to the Food and Drug Administration (FDA) an analysis that included events in the 6 months after trial termination, an analysis that produced a diminution of the dangers associated with salmeterol. The FDA eventually obtained the correct analysis [24]. The correct SMART analysis was finally published in January 2006 in a specialty journal, *CHEST* [25].

In another more recent example, Schering-Plough delayed, for almost 2 years, publication of a study of more than 700 patients that investigated a combination drug, ezetimibe and simvastatin vs. simvastatin alone, for improving lipid profiles and preventing atherosclerosis [26]. A review of submissions to the FDA in 2001 and 2002 found that many trials were still not published 5 years after FDA approval [27]. These examples of lag time bias demonstrate the need for avoiding excessive enthusiasm about early findings with new agents.

6. When to rate down for publication bias—industry influence

In general, review authors and guideline developers should consider rating down for likelihood of publication bias when the evidence consists of a number of small studies [17–21]. The inclination to rate down for publication bias should increase if most of those small studies are industry sponsored or likely to be industry sponsored (or if the investigators share another conflict of interest) [14,23,28].

Table 1. Publication bias

Phases of research publication	Actions contributing to or resulting in bias
Preliminary and pilot studies	Small studies more likely to be "negative" (e.g., those with discarded or failed hypotheses) remain unpublished; companies classify some as proprietary information
Report completion	Authors decide that reporting a "negative" study is uninteresting; and do not invest the time and effort required for submission
Journal selection	Authors decide to submit the "negative" report to a nonindexed, non-English, or limited-circulation journal
Editorial consideration	Editor decides that the "negative" study does not warrant peer review and rejects manuscript
Peer review	Peer reviewers conclude that the "negative" study does not contribute to the field and recommend rejecting the manuscript. Author gives up or moves to lower impact journal. Publication delayed
Author revision and resubmission	Author of rejected manuscript decides to forgo the submission of the "negative" study or to submit it again at a later time to another journal (see "journal selection," above).
Report publication	Journal delays the publication of the "negative" study
	Proprietary interests lead to report getting submitted to, and accepted by, different journals

An investigation of 74 antidepressant trials with a mean sample size of fewer than 200 patients submitted to the FDA illustrates the paradigmatic situation [28]. Of the 38 studies viewed as positive by the FDA, 37 were published. Of the 36 studies viewed as negative by the FDA, 14 were published. Publication bias of this magnitude can seriously bias effect estimates.

Additional criteria for suspicion of publication bias include a relatively recent RCT or set of RCTs addressing a novel therapy and systematic review authors' failure to conduct a comprehensive search (including a search for unpublished studies).

7. Using study results to estimate the likelihood of publication bias

Another criterion for publication bias is the pattern of study results. Suspicion may increase if visual inspection demonstrates an asymmetrical (Fig. 1b) rather than a symmetrical (Fig. 1a) funnel plot or if statistical tests of asymmetry are positive [29,30]. Although funnel plots may be helpful, review authors and guideline developers should bear in mind that visual assessment of funnel plots is distressingly prone to error [31,32]. Enhancements of funnel plots may (or may not) help to improve reproducibility and validity associated with their use [33].

Statisticians have developed quantitative methods that rely on the same principles [29,30]. Other statisticians have, however, questioned their appropriateness [7,34–36].

Furthermore, systematic review and guideline authors should bear in mind that even if they find convincing evidence of asymmetry, publication bias is not the only explanation. For instance, if smaller studies suffer from greater study limitations, they may yield biased overestimates of effects. Another explanation would be that, because of a more restrictive (and thus responsive) population, or a more careful administration of the intervention, the effect may actually be larger in the small studies.

A second set of tests, referred to as "trim and fill," tries to impute missing information and address its impact. Such tests begin by removing small "positive" studies that do not have a "negative" study counterpart. This leaves a symmetric funnel plot that allows calculation of a putative true effect. The investigators then replace the "positive" studies they have removed and add hypothetical studies that mirror these "positive" studies to create a symmetrical funnel plot that retains the new pooled effect estimate [21]. The same alternative explanations to asymmetry that we have noted for funnel plots apply here, and the imputation of new missing studies represents a daring assumption that would leave many uncomfortable.

Another set of tests estimates whether there are differential chances of publication based on the level of statistical significance [37,38]. These tests are well established in the educational and psychology literature but, probably

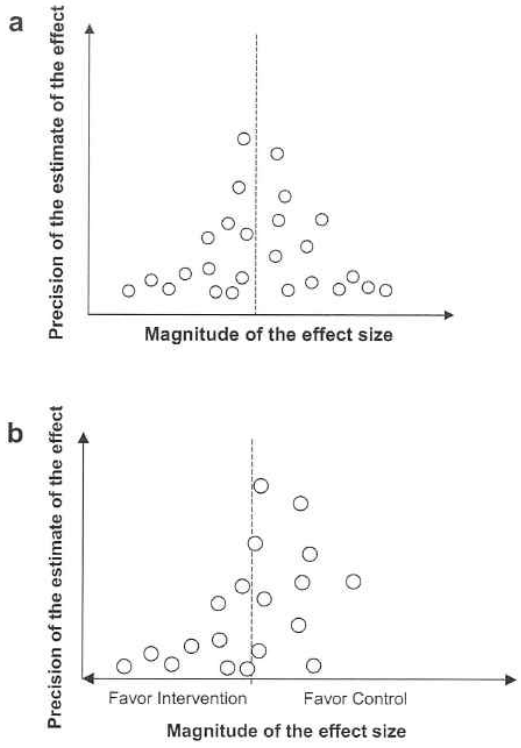


Fig. 1. (a) Funnel plot. The circles represent the point estimates of the trials. The pattern of distribution resembles an inverted funnel. Larger studies tend to be closer to the pooled estimate (the dashed line). In this case, the effect sizes of the smaller studies are more or less symmetrically distributed around the pooled estimate. (b) Publication bias. This funnel plot shows that the smaller studies are not symmetrically distributed around either the point estimate (dominated by the larger trials) or the results of the larger trials themselves. The trials expected in the bottom right quadrant are missing. One possible explanation for this set of results is publication bias—an overestimate of the treatment effect relative to the underlying truth.

because of their computational difficulty and complex assumptions, are uncommonly used in the medical sciences.

Finally, a set of tests examines whether evidence changes over time. Recursive cumulative meta-analysis [39] performs a meta-analysis at the end of each year for trials ordered chronologically and notes changes in the summary effect. Continuously diminishing effects strongly suggests time lag bias. Another test examines whether the number of statistically significant results is larger than what would be expected under plausible assumptions [40].

In summary, each of the approaches to using available data to provide insight into the likelihood of publication bias may be useful but has limitations. Concordant results of using more than one approach may strengthen inferences regarding publication bias.

More compelling than any of these theoretical exercises is authors' success in obtaining the results of some unpublished studies and demonstrating that the published and

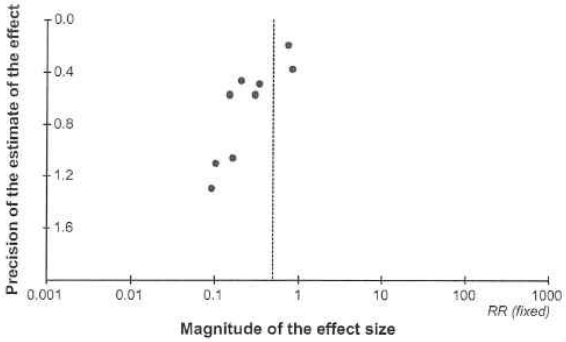


Fig. 2. Funnel plot of studies of flavonoids for ameliorating symptoms in patients with hemorrhoids [48]. RR, risk ratio.

unpublished data show different results. In these circumstances, the possibility of publication bias looms large. For instance, a systematic review found that including unpublished studies of the use of quinine for the treatment of leg cramps decreased the estimated effect size by a factor of two [41]. Unfortunately, obtaining the unpublished studies is not easy (although reliance on FDA submissions [or submissions to other regulatory agencies], as demonstrated in a number of examples we cited, can be very effective). On the other hand, reassurance may come from a systematic review that has succeeded in gaining industry cooperation and states that all trials have been revealed [42].

Prospective registration of all RCTs at inception and before their results become available enables review authors (and those using systematic reviews) to know when relevant trials have been conducted so that they can ask the responsible investigators for the relevant study data [43,44]. Mandatory registration of RCTs may be the only reliable method of addressing publication bias, and it is becoming increasingly common [45]. Consequently, searching clinical trial registers is becoming increasingly valuable and should be considered by review authors and those using systematic reviews when assessing the risk of publication bias. There is currently no initiative for registration of observational studies, leaving them, for the foreseeable future, open to publication bias.

8. Publication bias in observational studies

The risk of publication bias is probably larger for observational studies than for RCTs [3,32], particularly small observational studies and studies conducted on data collected automatically (e.g., in the electronic medical record or in a diabetes registry) or data collected for a previous study. In these instances, it is difficult for the reviewer to know if the observational studies that appear in the literature represent all or a fraction of the studies conducted, and whether the analyses in them represent all or a fraction of those conducted. In these instances, reviewers may consider the risk of publication bias as substantial [46,47].

9. Rating down for publication bias—an example

A systematic review of flavonoids in patients with hemorrhoids provides an example of a body of evidence in which rating down for publication bias is likely appropriate [48]. All trials, which ranged in size from 40 to 234 patients—with most around 100—were industry sponsored. Furthermore, the funnel plot suggests the possibility of publication bias (Fig. 2).

10. Acknowledging the difficulties in assessing the likelihood of publication bias

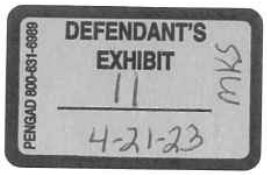
Unfortunately, it is very difficult to be confident that publication bias is absent, and almost equally difficult to know where to place the threshold and rate down for its likely presence. Recognizing these challenges, the terms GRADE suggests using in GRADE evidence profiles for publication bias are "undetected" and "strongly suspected." Acknowledging the uncertainty, GRADE suggests rating down a maximum of one level (rather than two) for suspicion of publication bias. Nevertheless, the examples cited herein suggest that publication bias is likely frequent, particularly in industry-funded studies. This suggests the wisdom of caution in the face of early results, particularly with small sample size and number of events.

References

[1] Montori V, Ioannidis J, Guyatt G. Reporting bias. In: Guyatt G, et al, editors. Users' guides to the medical literature: a manual for evidence-based clinical practice. New York, NY: McGraw-Hill; 2008.
[2] Hopewell S, Louden K, Clarke M, Oxman D, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev 2009; MR000006.
[3] Dickersin K, Min YI, Meinert CL. Factors influencing publication of research results. Follow-up of applications submitted to two institutional review boards. JAMA 1992;267:374–8.
[4] Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. BMJ 1997;315:640–5.
[5] Bardy AH. Bias in reporting clinical trials. Br J Clin Pharmacol 1998;46:147–50.
[6] Egger M, Smith GD. Bias in location and selection of studies. BMJ 1998;316:61–6.
[7] Song F, Eastwood A, Gilbody S, Duley L, Sutton A. Publication and related biases. Health Technol Assess 2000;4:1–115.
[8] Hopewell S, Clarke M, Steward L, Tierney J. Time to publication for results of clinical trials. Cochrane Database Syst Rev 2008; MR000011.
[9] Egger M, Zellweger-Zahner T, Schneider M, Junker C, Lengeler C, Antes G. Language bias in randomised controlled trials published in English and German. Lancet 1997;350:326–9.
[10] Juni P, Hollenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol 2002;31:115–23.

[11] Hopewell S, McDonald S, Clarke M, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev 2007; MR000010.
[12]. Rennie D. Fair conduct and fair reporting of clinical trials. JAMA 1999;282:1766–8.
[13] Tramer MR, Reynolds D, Moore R, McQuay H. Impact of covert duplicate publication on meta-analysis: a case study. BMJ 1997;315:635–40.
[14] Melander H, Ahlqvist-Rastad J, Meijer G, Beerman B. Evidence b(i)ased medicine—selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. BMJ 2003;326:1171–3.
[15] von Elm E, et al. Different patterns of duplicate publication: an analysis of articles used in systematic reviews. JAMA 2004;291:974–80.
[16] Vaitkus PT, Brar C. N-acetylcysteine in the prevention of contrast-induced nephropathy: publication bias perpetuated by meta-analyses. Am Heart J 2007;153:275–80.
[17] Begg CB, Berlin JA. Publication bias and dissemination of clinical research. J Natl Cancer Inst 1989;81:107–15.
[18] Egger M, et al. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629–34.
[19] Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.
[20] Cappelleri JC, et al. Large trials vs meta-analysis of smaller trials: how do their results compare? JAMA 1996;276:1332–8.
[21] Sutton AJ, et al. Empirical assessment of effect of publication bias on meta-analyses. BMJ 2000;320:1574–7.
[22] Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. JAMA 1998;279:281–6.
[23] Lexchin J, et al. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. BMJ 2003;326:1167–70.
[24] Lurie P, Wolfe S. Misleading data analyses in salmeterol (SMART) study. Lancet 2005;366:1261–2.
[25] Nelson HS, et al. The Salmeterol Multicenter Asthma Research Trial: a comparison of usual pharmacotherapy for asthma or usual pharmacotherapy plus salmeterol. Chest 2006;129:15–26.
[26] Mitka M. Controversies surround heart drug study: questions about vytorin and trial sponsors' conduct. JAMA 2008;299:885–7.
[27] Rising K, Bacchetti P, Bero L. Reporting bias in drug trials submitted to the Food and Drug Administration: review of publication and presentation. PLoS Med 2008;5:e217. discussion e217.
[28] Turner EH, et al. Selective publication of antidepressant trials and its influence on apparent efficacy. N Engl J Med 2008;358:252–60.
[29] Begg C, Berlin J. Publication bias: a problem in interpreting medical data. J R Statist Soc A 1988;151:419–63.
[30] Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50:1088–101.
[31] Terrin N, Schmid CH, Lau J. In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. J Clin Epidemiol 2005;58:894–901.
[32] Lau J, et al. The case of the misleading funnel plot. BMJ 2006; 333:597–600.
[33] Peters JL, et al. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol 2008;61:991–6.
[34] Irwig L, Macaskill P, Berry G, Glasziou P. Bias in meta-analysis detected by a simple, graphical test. Graphical test is itself biased. discussion 470–471. BMJ 1998;316:470.
[35] Stuck A, Rubenstein L, Wieland D. Bias in meta-analysis detected by a simple, graphical test. Asymmetry detected in funnel plot was probably due to true heterogeneity. BMJ 1998;316:469.
[36] Seagroatt V, Stratton I. Bias in meta-analysis detected by a simple, graphical test. Test had 10% positive rate. discussion 470–471. BMJ 1998;316:470.
[37] Hedges L, Vevea J. Estimating effect size under publication bias: small sample properties and robustness of a random effects selection model. J Educ Behav Stat 1996;21:299–333.
[38] Vevea J, Hedges L. A general linear model for estimating effect size in the presence of publication bias. Psychometrika 1995;60: 419–35.
[39] Ioannidis JP, Contopoulos-Ioannidis DG, Lau J. Recursive cumulative meta-analysis: a diagnostic for the evolution of total randomized evidence from group and individual patient data. J Clin Epidemiol 1999;52:281–91.
[40] Pan Z, Trikalinos T, Kavvoura F, Lau J, Ioannidis J. Local literature bias in genetic epidemiology: an empirical evaluation of the Chinese literature. PLoS Med 2011;2(12):e334.
[41] Man-Son-Hing M, Wells G, Lau A. Quinine for nocturnal leg cramps: a meta-analysis including unpublished data. J Gen Intern Med 1998;13:600–6.
[42] Cranney A, Wells G, Wilan A, Griffith L, Zytaruk N, Robinson V, et al. Meta-analyses of therapies for postmenopausal osteoporosis. II. Meta-analysis of alendronate for the treatment of postmenopausal women. Endocr Rev 2002;23:508–16.
[43] DeAngelis CD, Drazen J, Frizelle F, Haug C, Hoey J, Horton R, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. JAMA 2004;292:1363–4.
[44] Gulmezoglu AM, Pang T, Horton R, Dickersin K. WHO facilitates international collaboration in setting standards for clinical trial registration. Lancet 2005;365:1829–31.
[45] Laine C, Horton R, DeAngelis CD, Drazen JM, Frizelle FA, Godlee F, et al. Clinical trial registration—looking back and moving ahead. N Engl J Med 2007;356:2734–6.
[46] Easterbrook PJ, Gopalan R, Berlin J, Matthews D. Publication bias in clinical research. Lancet 1991;337:867–72.
[47] Altman DG. Systematic reviews of evaluations of prognostic variables. BMJ 2001;323:224–8.
[48] Alonso-Coello P, Zhou Q, Martinez-Zapata M, Mills E, Heels-Ansdell D, Johansen J, Guyatt G. Meta-analysis of flavonoids for the treatment of haemorrhoids. Br J Surg 2006;93:909–20.

i An update to this article is included at the end



ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1283–1293

Journal of Clinical Epidemiology

GRADE guidelines 6. Rating the quality of evidence—imprecision

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

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

Accepted 5 January 2011; Published online 11 August 2011

Abstract

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

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

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

1. Introduction

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

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

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

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

0895-4356/$ - see front matter
doi: 10.1016/j.jclinepi.2011.01.012

Key Points

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

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

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

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

2. Criteria for imprecision differ for guidelines and systematic reviews

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

3. Confidence intervals capture the extent of imprecision—mostly

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

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

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

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

Further, assume that the intervention is a drug with no serious adverse effects, minimal inconvenience, and modest cost. Under these circumstances, even a small effect would warrant a strong recommendation. For instance, we may strongly recommend the intervention were it to reduce strokes by as little as 0.5% (vertical middle line in Fig. 1)—an NNT of 200. The entire CI (0.6% to 2.0%) around the effect on stroke reduction lies to the left of the clinical decision threshold of 0.5% and therefore excludes a benefit smaller than the threshold. We can therefore conclude that

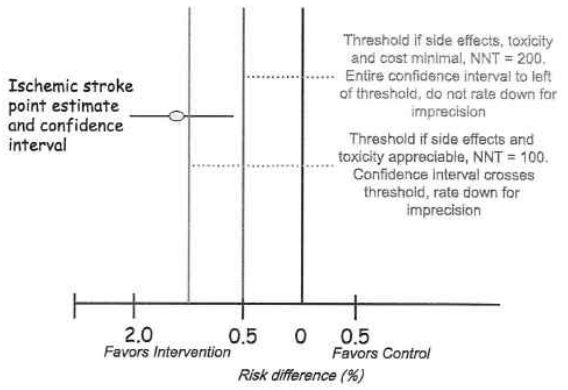


Fig. 1. Rating down for imprecision in guidelines: thresholds are key.

Box 1 The impact of undesirable consequences on precision

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

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

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

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

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

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

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

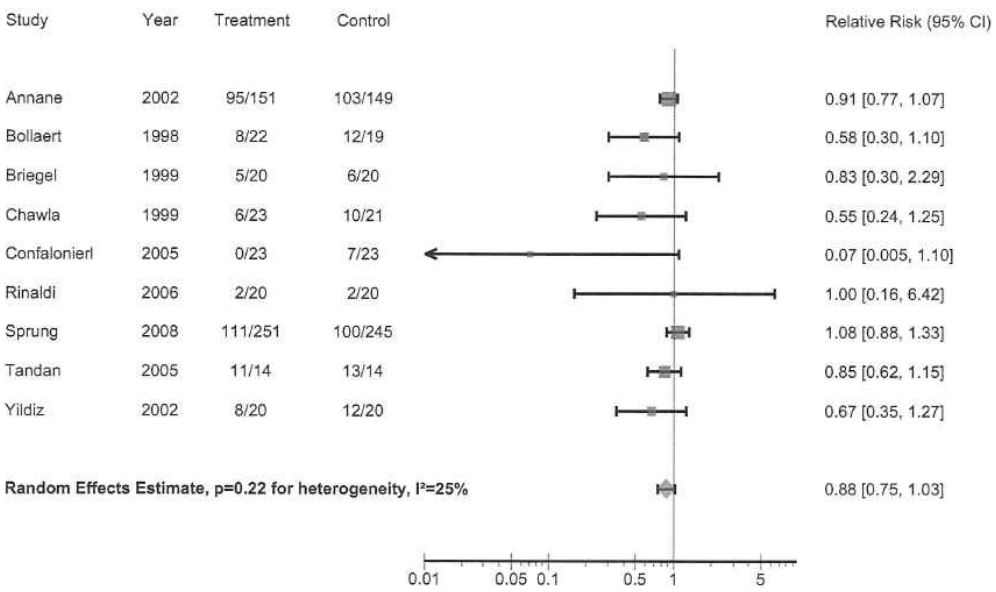


Fig. 2. Corticosteroids to reduce hospital mortality in septic shock.

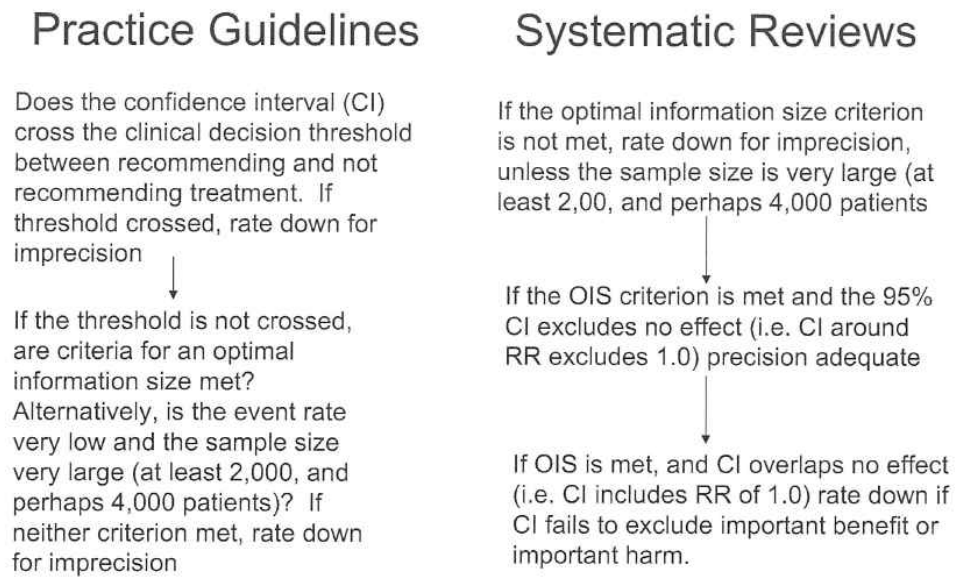


Fig. 3. Deciding whether to rate down for imprecision in guidelines and systematic reviews of binary variables.

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

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

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

6. Confidence intervals can be misleading because of fragility

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

7. The danger of initial trials with impressive positive results

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

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

Box 3 An example of fragility

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

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

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

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

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

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

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

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

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

Box 4 presents examples of application of the OIS.

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

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

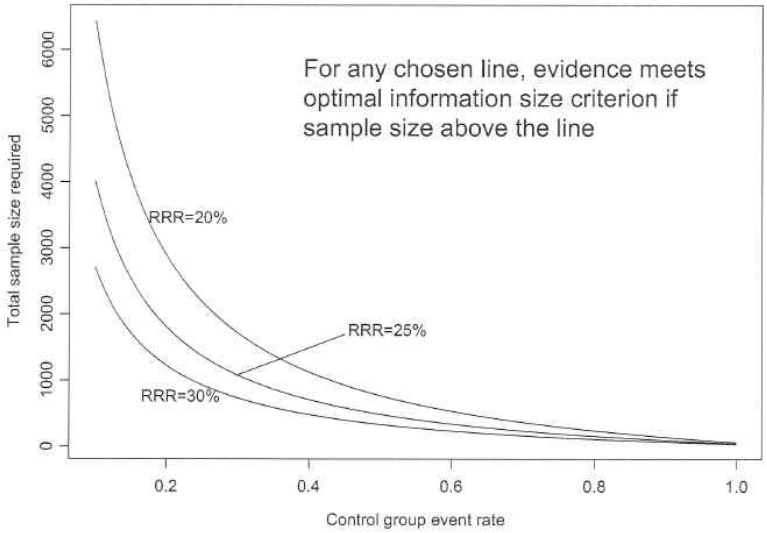


Fig. 4. Optimal information size given α of 0.05 and β of 0.2 for varying control event rates and relative risks.

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

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

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

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

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

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

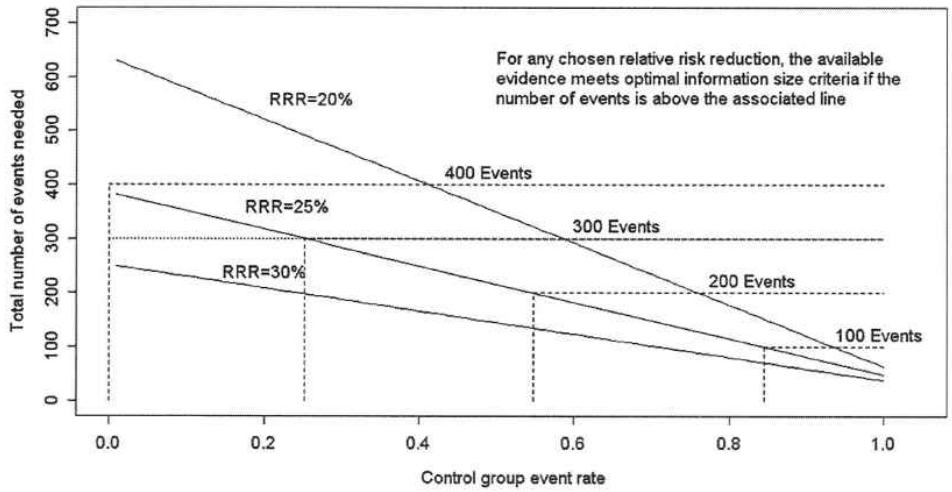


Fig. 5. Optimal information size presented as number of events given α of 0.05 and β of 0.20 for varying control event rates and RRR of 20%, 25% and 30%. *Abbreviation*: RRR, relative risk reduction.

Table 1. Optimal information size implications from Fig. 5

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

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

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

There were, however, only 34 SAEs in over 2,700 patients, corresponding to event rates of 1.2 and 1.3% in the two groups. Of these events, 2 were deaths, 2 severe anemias, and the remainder febrile seizures and elevated liver function. In absolute terms, the difference between groups is 1 event per thousand patients with a CI from 6 in 1,000 fewer to 14 in 1,000 more. Particularly considering that very few of these SAEs were associated with long-term morbidity,

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

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

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

focusing on the CI around absolute versus relative effects would lead one to reject rating down quality two levels for imprecision, and possibly not rate down for imprecision at all. Box 5 presents a second example of this issue.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

13. Rating down two levels for imprecision

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

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

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

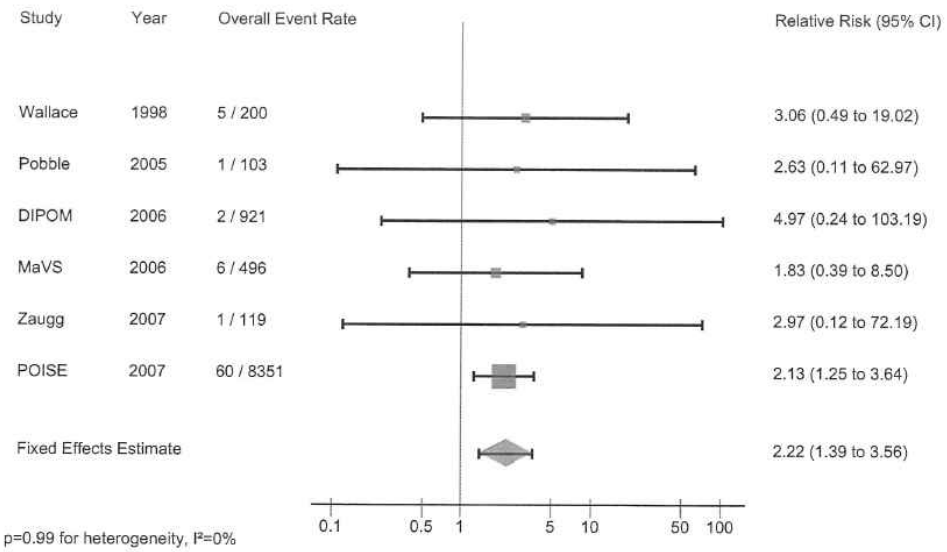


Fig. 6. Meta-analysis of beta blockers in noncardiac surgery: outcome and stroke.

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

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

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

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

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

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

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

15. Standards for adequate precision in guidelines addressing continuous variables

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

Box 7 Dealing with close call decisions

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

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

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

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

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

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

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

16. Conclusion

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

References

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

[12] Yusuf S, Collins R, MacMahon S, et al. Effect of intravenous nitrates on mortality in acute myocardial infarction: an overview of the randomised trials. Lancet 1988;1:1088–92.
[13] GISSI-3: effects of lisinopril and transdermal glyceryl trinitrate singly and together on 6-week mortality and ventricular function after acute myocardial infarction. Gruppo Italiano per lo Studio della Sopravvivenza nell'infarto Miocardico. Lancet 1994;343:1115–22.
[14] Imperiale TF, Petrulis AS. A meta-analysis of low-dose aspirin for the prevention of pregnancy-induced hypertensive disease. JAMA 1991;266:260–4.
[15] CLASP: a randomised trial of low-dose aspirin for the prevention and treatment of pre-eclampsia among 9364 pregnant women. CLASP (Collaborative Low-dose Aspirin Study in Pregnancy) Collaborative Group. Lancet 1994;343:619–29.
[16] Human albumin administration in critically ill patients: systematic review of randomised controlled trials. Cochrane Injuries Group Albumin Reviewers. BMJ 1998;317:235–40.
[17] Finfer S, Bellomo R, Boyce N, et al. A comparison of albumin and saline for fluid resuscitation in the intensive care unit. N Engl J Med 2004;350:2247–56.
[18] Trikalinos TA, Churchill R, Ferri M, et al. Effect sizes in cumulative meta-analyses of mental health randomized trials evolved over time. J Clin Epidemiol 2004;57:1124–30.
[19] Pogue JM, Yusuf S. Cumulating evidence from randomized trials: utilizing sequential monitoring boundaries for cumulative meta-analysis. Control Clin Trials 1997;18:580–93. discussion 661–666.
[20] Alonso-Coello P, Zhou Q, Martinez-Zapata MJ, et al. Meta-analysis of flavonoids for the treatment of haemorrhoids. Br J Surg 2006; 93:909–20.
[21] Gafter-Gvili A, Fraser A, Paul M, et al. Meta-analysis: antibiotic prophylaxis reduces mortality in neutropenic patients. Ann Intern Med 2005;142:979–95.
[22] Murad MH, Flynn DN, Elamin MB, et al. Endarterectomy vs stenting for carotid artery stenosis: a systematic review and meta-analysis. J Vasc Surg 2008;48:487–93.
[23] Devereaux PJ, Yang H, Yusuf S, et al. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. Lancet 2008;371: 1839–47.
[24] Butterworth AD, Thomas AG, Akobeng AK. Probiotics for induction of remission in Crohn's disease. Cochrane Database of Systematic Reviews 2008; Issue 3. Art. No.: CD006634. doi:10.1002/14651858.CD006634.pub2.
[25] Quon BS, Gan WQ, Sin DD. Contemporary management of acute exacerbations of COPD: a systematic review and metaanalysis. Chest 2008;133:756–66.

Update

Journal of Clinical Epidemiology
Volume 137, Issue , September 2021, Page 265

DOI: https://doi.org/10.1016/j.jclinepi.2021.04.014

ELSEVIER

Check for updates

Journal of Clinical Epidemiology 137 (2021) 265–265

Journal of Clinical Epidemiology

CORRIGENDUM

Corrigendum to GRADE guidelines 6. Rating the quality of evidence-imprecision. J Clin Epidemiol 2011;64:1283–1293

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

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

The authors regret, in the above-mentioned article, we discovered an error related to standards for adequate precision in systematic reviews of continuous variables. The relevant paragraph reads as follows:

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

The sample size stated is incorrect: the correct number is 800 (400 per group).

The authors would like to apologise for any inconvenience caused. The authors are grateful to Mark Chatfield for pointing out this error.

DOI of original article: 10.1016/j.jclinepi.2011.01.012

https://doi.org/10.1016/j.jclinepi.2021.04.014

the choice of aspirin or heparin for venous thromboembolism prophylaxis among patients with operatively treated extremity fractures (or any pelvic or acetabular fracture), this is by far the largest trial to date and provides compelling evidence that a readily available, inexpensive drug, taken orally, is a viable alternative to an injectable pharmacologic prophylaxis.

Are there any caveats to this message? The trial shows several secondary outcomes that support the main conclusion of the trial, including a similar risk of pulmonary embolism in the two groups and, in terms of safety outcomes, no evidence of a difference in the incidence of bleeding events, which occurred in 13.72% of patients in the aspirin group and 14.27% in the low-molecular-weight–heparin group. However, in keeping with previous trials, the authors noted that deep-vein thrombosis was more frequent in patients who had received aspirin than in those who had received heparin (2.51% vs. 1.71%), although the absolute difference was small (0.80 percentage points). Although deep-vein thrombosis is clearly not as serious as a fatal pulmonary embolism, it is not an inconsequential problem. Post-thrombotic syndrome affects some people who have had a deep-vein thrombosis of the leg, and this condition can cause chronic pain and swelling.[9]

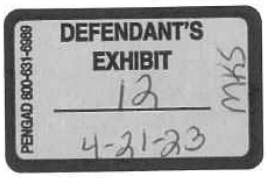
The findings in this trial clearly indicate that guidelines for the prevention of hospital-acquired venous thromboembolism will need to be rewritten to include the option of aspirin in patients with traumatic injuries. More work is needed to determine whether aspirin should also be considered for venous thromboembolism prophylaxis after other types of surgeries and for nonsurgical patients who have risk factors for venous thromboembolism.

Disclosure forms provided by the author are available with the full text of this editorial at NEJM.org.

From Oxford Trauma and Emergency Care, Nuffield Department of Orthopedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom.

1. National Institute for Health and Care Excellence. Venous thromboembolism in over 16s: reducing the risk of hospital-acquired deep vein thrombosis or pulmonary embolism. August 13, 2019 (https://www.nice.org.uk/guidance/ng89/chapter/Recommendations).
2. Hegsted D, Gritsiouk Y, Schlesinger P, Gardiner S, Gubler KD. Utility of the risk assessment profile for risk stratification of venous thrombotic events for trauma patients. Am J Surg 2013;205:517-20.
3. Geerts WH, Code KI, Jay RM, Chen E, Szalai JP. A prospective study of venous thromboembolism after major trauma. N Engl J Med 1994;331:1601-6.
4. Barrera LM, Perel P, Ker K, Cirocchi R, Farinella E, Morales Uribe CH. Thromboprophylaxis for trauma patients. Cochrane Database Syst Rev 2013;3:CD008303.
5. Rogers FB, Cipolle MD, Velmahos G, Rozycki G, Luchette FA. Practice management guidelines for the prevention of venous thromboembolism in trauma patients: the EAST Practice Management Guidelines Work Group. J Trauma 2002;53:142-64.
6. Colwell CW Jr, Pulido P, Hardwick ME, Morris BA. Patient compliance with outpatient prophylaxis: an observational study. Orthopedics 2005;28:143-7.
7. Horner D, Goodacre S, Pandor A, et al. Thromboprophylaxis in lower limb immobilisation after injury (TiLLI). Emerg Med J 2020;37:36-41.
8. Major Extremity Trauma Research Consortium (METRC). Aspirin or low-molecular-weight heparin for thromboprophylaxis after a fracture. N Engl J Med 2023;388:203-13.
9. Makedonov I, Kahn SR, Abdulrehman J, et al. Prevention of the postthrombotic syndrome with anticoagulation: a narrative review. Thromb Haemost 2022;122:1255-64.

DOI: 10.1056/NEJMe2214045



Growing Evidence and Remaining Questions in Adolescent Transgender Care

Annelou L.C. de Vries, M.D., Ph.D., and Sabine E. Hannema, M.D., Ph.D.

This week in the *Journal*, a much-awaited primary report from Chen et al.[1] on 2 years of gender-affirming hormones (GAH) in transgender adolescents appears. The approach to adolescent transgender care with early treatment with puberty blockers, and GAH in youth from 16 years of age, originated in the Netherlands ("the Dutch model") and became the dominant medical care model for transgender adolescents.[2] Especially over the past decade, marked increases in referrals but limited evidence as to long-term outcomes have led to controversies and debate regarding this approach. Indeed, some European countries are adapting their guidelines and re-

stricting access to care for transgender youth, and some states in the United States have introduced laws to ban such care.[3] Therefore, rigorous longitudinal outcome studies that provide evidence about whether this approach is effective and safe are needed.

The results of the current study — involving a large, multisite sample of 315 participants — provide such evidence. During 24 months of GAH treatment, participant-reported appearance congruence (alignment between gender identity and physical appearance), positive affect, and life satisfaction increased and depression and anxiety decreased. In addition, initial levels and rates of change in appearance congruence correlated with the psychosocial outcomes. These results corroborate the positive effects in several earlier studies of smaller samples of adolescents and add to the evidence base that GAH can have a positive effect on mental health.[4]

Yet the study leaves some concerns unanswered. Although overall psychological functioning in the study participants improved, there was substantial variation among participants; a considerable number still had depression, anxiety, or both at 24 months, and two died by suicide. The correlation between appearance congruence and various psychological-outcome variables suggests an important mediating role of GAH and consequent bodily changes. However, other possible determinants of outcomes were not reported, particularly the extent of mental health care provided throughout GAH treatment. To date, international guidelines for transgender adolescent care recommend a psychosocial assessment and involvement of mental health professionals in a multidisciplinary care model.[5] Whether participating centers in the current study followed that approach is unfortunately unclear. Future studies that compare outcomes with different care models are needed, preferably using similar measures.

In addition, some are concerned that young persons may not be capable of making decisions regarding medical treatments that have irreversible effects that they might regret later in life. In the 2-year study by Chen et al., 9 of 314 adolescents (2.9%) stopped GAH, but it is unclear whether they detransitioned or regretted their treatment or whether they stopped because they were satisfied with treatment-related changes. Despite concerns about detransitioning, few studies have provided data on the incidence of detransitioning, and available results are inconsistent. Although one U.S. study showed that 74% of adolescents who started GAH treatment were still receiving it 4 years later, 98% of 720 Dutch adolescents who began such therapy were receiving it after a median of 2.7 years (range, 0.0 to 20.0).[6,7] Similar studies in other centers, regions, and countries are necessary to learn whether the incidence of detransitioning differs between settings and what factors are associated with these differences. It will be especially important to evaluate outcomes in adolescents starting GAH before 16 years of age, the age limit in the initial Dutch protocol.[2]

Furthermore, although Chen et al. investigated relevant psychological and gender outcome measures (e.g., depression, appearance congruence, and life satisfaction), additional factors such as autism spectrum disorder and the quality of peer relations and family support are also of interest. Social support has been hypothesized as explaining why Dutch transgender adolescents have better psychological function than those in other countries.[8] Understanding additional factors that influence outcomes should help to determine which components of care and support other than GAH might improve the lives of transgender adolescents.

Finally, benefits of early medical intervention, including puberty suppression, need to be weighed against possible adverse effects — for example, with regard to bone and brain development and fertility. At present, studies involving young adults from the Dutch adolescent transgender cohort show that accrual of bone mineral decelerates during puberty suppression but increases during GAH treatment and also that adolescents' educational achievements are as expected given their pretreatment status, which is reassuring.[9,10] However, those results from a single Dutch center should be replicated and validated in other contexts, as in a sample followed in the current study.

Despite uncertainties that call for further study, current information shows that mental health improves with GAH, whereas withholding treatment may lead to increased gender dysphoria and adversely affect psychological functioning. The study by Chen et al. adds to the

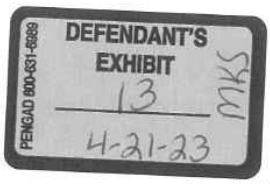
evidence of the effectiveness of the current care model that includes hormonal treatment for transgender adolescents.

Disclosure forms provided by the authors are available with the full text of this editorial at NEJM.org.

From the Departments of Child and Adolescent Psychiatry (A.L.C.V.) and Pediatrics (S.E.H.), Center of Expertise on Gender Dysphoria, Amsterdam University Medical Centers, Location Vrije Universiteit, Amsterdam.

1. Chen D, Berona J, Chan Y-M, et al. Psychosocial functioning in transgender youth after 2 years of hormones. N Engl J Med 2023;388:240-50.
2. Delemarre-van de Waal HA, Cohen Kettenis PT. Clinical management of gender identity disorder in adolescents: a protocol on psychological and paediatric endocrinology aspects. Eur J Endocrinol 2006;155:Suppl 1:S131-S137.
3. Cass Review. Independent review of gender identity services for children and young people: interim report. February 2022 (https://cass.independent-review.uk/publications/interim-report/).
4. Kuper LE, Stewart S, Preston S, Lau M, Lopez X. Body dissatisfaction and mental health outcomes of youth on gender-affirming hormone therapy. Pediatrics 2020;145(4):e20193006.
5. Coleman E, Radix AE, Bouman WP, et al. Standards of care for the health of transgender and gender diverse people, version 8. Int J Transgend Health 2022;23:Suppl 1:S1-S259.
6. Roberts CM, Klein DA, Adirim TA, Schvey NA, Hisle-Gorman E. Continuation of gender-affirming hormones among transgender adolescents and adults. J Clin Endocrinol Metab 2022;107(9):e3937-e3943.
7. van der Loos MATC, Hannema SE, Klink DT, den Heijer M, Wiepjes CM. Continuation of gender-affirming hormones in transgender people starting puberty suppression in adolescence: a cohort study in the Netherlands. Lancet Child Adolesc Health 2022;6:869-75.
8. de Graaf NM, Steensma TD, Carmichael P, et al. Suicidality in clinic-referred transgender adolescents. Eur Child Adolesc Psychiatry 2022;31:67-83.
9. Schagen SEE, Schagen SEE, Wouters FM, Cohen-Kettenis PT, Gooren LJ, Hannema SE. Bone development in transgender adolescents treated with GnRH analogues and subsequent gender-affirming hormones. J Clin Endocrinol Metab 2020;105(12):e4252-e4263.
10. Arnoldussen M, Hooijman EC, Kreukels BP, de Vries AL. Association between pre-treatment IQ and educational achievement after gender-affirming treatment including puberty suppression in transgender adolescents. Clin Child Psychol Psychiatry 2022;27:1069-76.

DOI: 10.1056/NEJMe2216191



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
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.

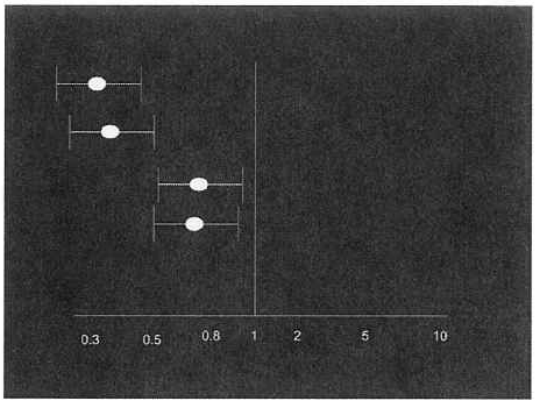


Fig. 2. Substantial heterogeneity, but of questionable importance.

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.

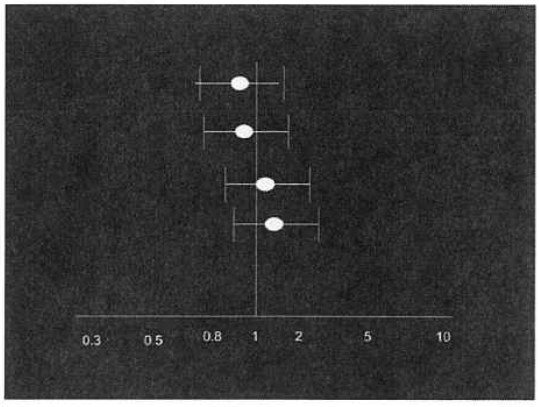


Fig. 1. Differences in direction, but minimal heterogeneity.

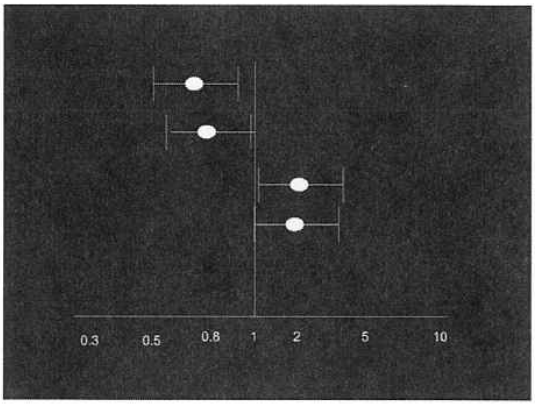


Fig. 3. Substantial heterogeneity, of unequivocal importance.

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.

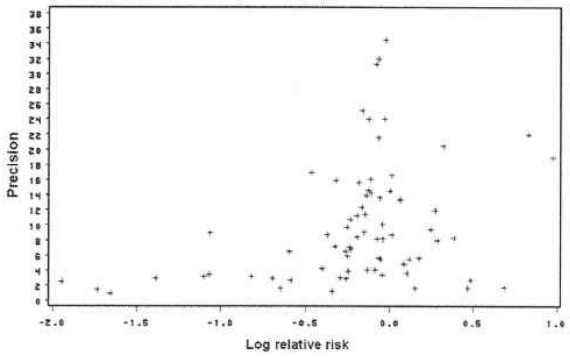


Fig. 4. Funnel plot for all-cause mortality, United States vs. Canada. Negative values favor Canada, positive values, United States.

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

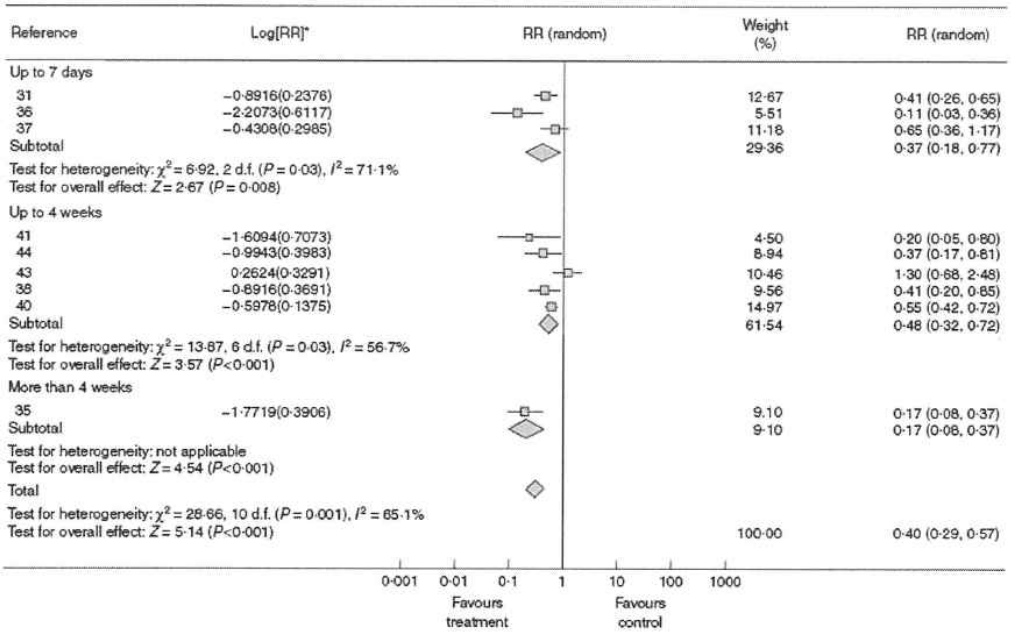


Fig. 5. Results of a systematic review of flavanoids for treatment of hemorrhoids: relative risks of failure to improve.

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.

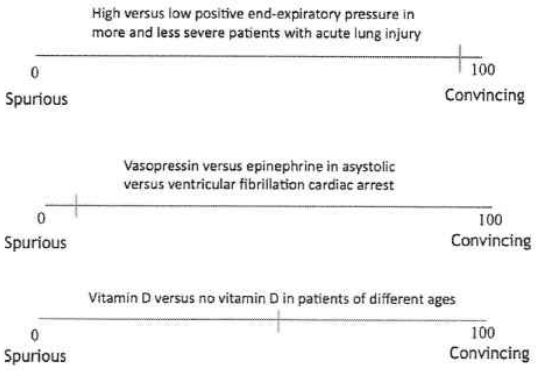


Fig. 6. Credibility of subgroup analyses from three systematic reviews.

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

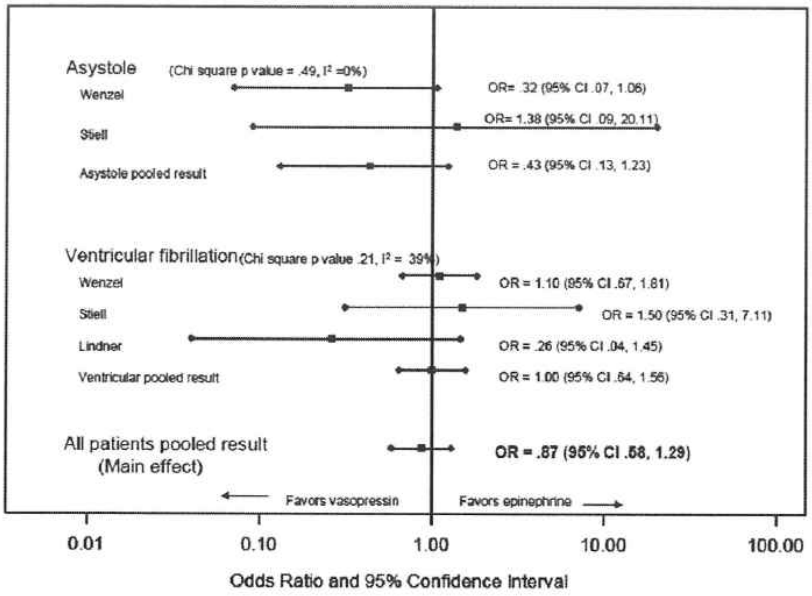


Fig. 7. Vasopressin vs. epinephrine in cardiac arrest.

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.

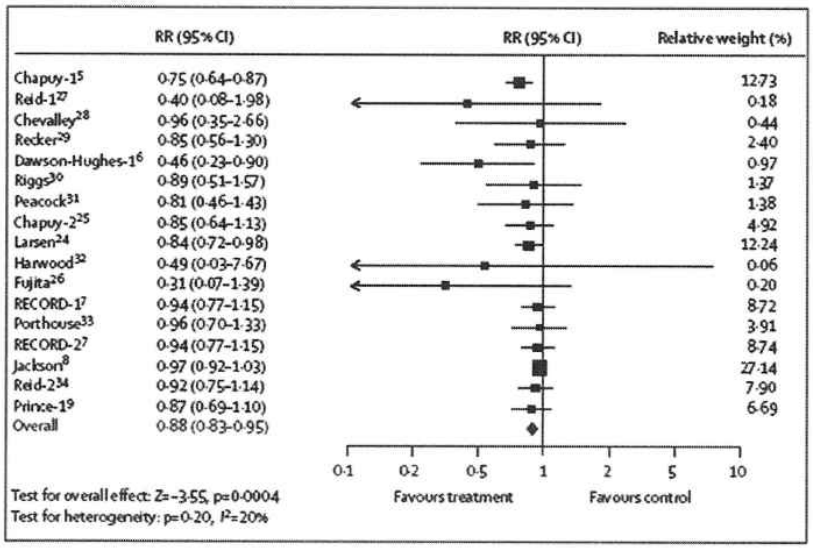


Fig. 8. Fracture reduction with calcium in patients older than 50 yr.

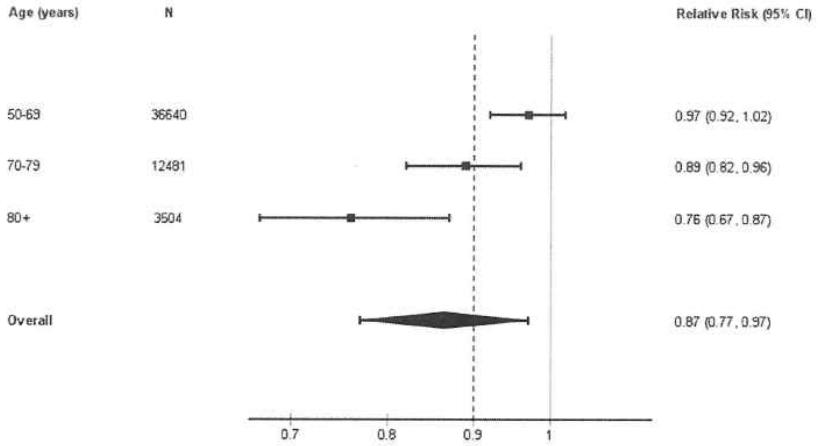


Fig. 9. Apparent fracture reduction with calcium in patients in three age ranges.

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.

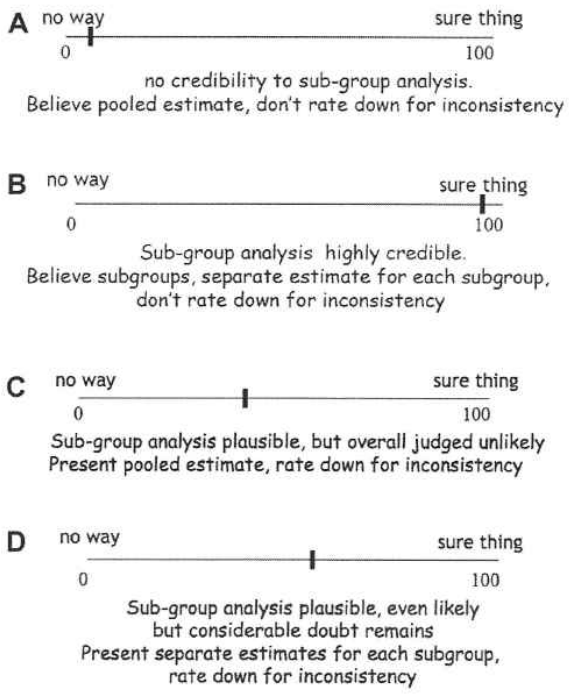


Fig. 10. Interpretation of subgroup analyses of varying credibility.

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.

ELSEVIER

Journal of Clinical Epidemiology 64 (2011) 1303–1310

Journal of Clinical Epidemiology

GRADE guidelines: 8. Rating the quality of evidence—indirectness

Gordon H. Guyatt[a,b,*], Andrew D. Oxman[c], Regina Kunz[d,e], James Woodcock[f], Jan Brozek[a], Mark Helfand[g], Pablo Alonso-Coello[h], Yngve Falck-Ytter[i,j], Roman Jaeschke[b], Gunn Vist[c], Elie A. Akl[k], Piet N. Post[l], Susan Norris[m], Joerg Meerpohl[n,o], Vijay K. Shukla[p], Mona Nasser[q], 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]*Academy of Swiss Insurance Medicine (asim), University Hospital Basel Petergraben 4, CH-4031, Basel, Switzerland*
[e]*Basel Institute of Clinical Epidemiology, University Hospital Basel Hebelstrasse 10, 4031 Basel, Switzerland*
[f]*London School of Hygiene and Tropical Medicine, London, United Kingdom*
[g]*Oregon Evidence-Based Practice Center, Oregon Health and Science University, Portland VA Medical Center, Portland, OR, USA*
[h]*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*
[i]*Division of Gastroenterology, Case and VA Medical Center, Case Western Reserve University, Cleveland, OH 44106, USA*
[j]*University of Oxford, Oxford, United Kingdom*
[k]*Department of Medicine, State University of New York at Buffalo, NY, USA*
[l]*Dutch Institute for Healthcare Improvement CBO, Utrecht, The Netherlands*
[m]*Department of Medical Informatics and Clinical Epidemiology, Oregon Health and Science University, Portland, OR 97239-3098, 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*
[p]*Canadian Agency for Drugs and Technology in Health (CADTH), Ottawa K1S 5S8, Canada*
[q]*Department of health information, Institute for Quality and Efficiency in Health care (IQWiG), Cologne, Germany*

Accepted 18 April 2011; Published online 30 July 2011

Abstract

Direct evidence comes from research that directly compares the interventions in which we are interested when applied to the populations in which we are interested and measures outcomes important to patients. Evidence can be indirect in one of four ways. First, patients may differ from those of interest (the term applicability is often used for this form of indirectness). Secondly, the intervention tested may differ from the intervention of interest. Decisions regarding indirectness of patients and interventions depend on an understanding of whether biological or social factors are sufficiently different that one might expect substantial differences in the magnitude of effect.

Thirdly, outcomes may differ from those of primary interest—for instance, surrogate outcomes that are not themselves important, but measured in the presumption that changes in the surrogate reflect changes in an outcome important to patients.

A fourth type of indirectness, conceptually different from the first three, occurs when clinicians must choose between interventions that have not been tested in head-to-head comparisons. Making comparisons between treatments under these circumstances requires specific statistical methods and will be rated down in quality one or two levels depending on the extent of differences between the patient populations, co-interventions, measurements of the outcome, and the methods of the trials of the candidate interventions.

Keywords: GRADE; Quality of evidence; Indirectness; Indirect comparisons; Applicability; Generalizability

[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.2011.04.014

Key points

- Quality of evidence (our confidence in estimates of effect) may decrease when substantial differences exist between the population, the intervention, or the outcomes measured in relevant research studies and those under consideration in a guideline or systematic review.
- Quality of evidence decreases if head-to-head comparisons are unavailable. Such instances require falling back on indirect comparisons in which, for example, we make inferences about the relative effect of two interventions on the basis of their comparison not with one another, but with a third or control condition.

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, imprecision, and inconsistency. In this article, we deal with another potential problem: indirectness.

2. Four types of indirectness

We are more confident in the results when we have direct evidence. By direct evidence, we mean research that directly compares the interventions in which we are interested delivered to the populations in which we are interested and measures the outcomes important to patients. Thus, we can have concerns about indirectness when the population, intervention, or outcomes differ from those in which we are interested (Table 1). A fourth, different type of indirectness, occurs when there are no head-to-head comparisons between the alternative management strategies under comparison (Table 1). Indirectness of outcomes and indirect comparisons are equally relevant to systematic reviews and practice guidelines; indirectness related to populations and interventions (sometimes referred to as applicability) is more relevant to guidelines.

2.1. Indirectness: differences in population (applicability)

The first type of indirectness includes differences between the population of interest and those who have participated in relevant studies. Systematic reviews will include only patients who meet the reviews' eligibility criteria; thus, in a sense, evidence regarding patients is direct by definition.

There may, however, be exceptions. For example, a systematic review might have an a priori hypothesis that a drug would have different effects in children than in adults based on what is known about the mechanism of action. If no studies were found that tested the drug in children, the review authors might conclude that the effects in children were less certain than in adults, based on the indirectness of the evidence for children.

Differences between the population of interest and those in studies are a common problem for guideline developers who will seek the best evidence relevant to their question. For instance, a World Health Organization guideline panel addressed the treatment of infection with avian influenza A virus but needed to use evidence from seasonal influenza (Table 1; Box 1) [1].

Less extreme differences in patients (or the conditions from which they suffer) would lead to rating down only one level, or even no rating down whatsoever. Because randomized trial eligibility criteria often exclude patients with comorbidity, as guideline developers begin to address issues of multiple coexisting conditions (patients with, for instance, heart failure and asthma) they will often need to consider issues of indirectness. Some population differences may be partly addressed by subgroup analyses within the trials or reviews that check the robustness of the results across population factors such as age, gender, or disease severity. For example, pooled analyses of large-scale trials of statins show highly consistent relative risk (RR) reductions across a wide variety of subpopulations.

In general, one should not rate down for population differences unless one has compelling reason to think that the biology in the population of interest is so different from that

Table 1. Evidence is lower quality if comparisons are indirect

Question of interest	Source of indirectness
Oseltamivir for prophylaxis of avian flu caused by influenza A virus	*Differences in population*: Randomized trials of oseltamivir are available for seasonal influenza, but not for avian flu
Colonoscopic screening for prevention of colon cancer mortality	*Differences in intervention*: Randomized trials of fecal occult blood screening provide indirect evidence bearing on the potential effectiveness of colonoscopy
Sevelamer- vs. calcium-based phosphate binders in chronic renal failure	*Differences in outcome*: Reducing the calcium-phosphate load is hypothesized to reduce vascular calcification, which is hypothesized to reduce vascular events
Choice of antidepressant	*Indirect comparison*: Some antidepressants have been compared directly with others, but many have not

Box 1 Indirectness of population: avian influenza

High-quality randomized trials have demonstrated the effectiveness of antiviral treatment for seasonal influenza. The panel judged that the biology of seasonal influenza was sufficiently different from that of avian influenza (that is, the avian influenza organism may be far less responsive to the available antiviral agents than seasonal influenza) that the evidence required rating down by two levels for indirectness.

of the population tested that the magnitude of effect will differ substantially. Most often, this will not be the case. Note that we are referring here to consistency in RR: differences in baseline risk or control event rate in subpopulations will, on many occasions, lead to difference in absolute effect between subgroups.

The above discussion refers to different human populations, but sometimes the only evidence will be from animal studies, such as rats or primates. In general, we would rate such evidence down two levels for indirectness. Animal studies may, however, provide an important indication of drug toxicity. Although toxicity data from animals does not reliably predict toxicity in humans, evidence of animal toxicity should engender caution in recommendations.

Another type of nonhuman study may generate high-quality evidence. Consider laboratory evidence of change in resistance patterns of bacteria to antimicrobial agents (e.g., the emergence of methicillin-resistant staphylococcus aureus—MRSA). These laboratory findings may constitute high-quality evidence for the superiority of antibiotics to which MRSA is sensitive vs. methicillin as the initial treatment of suspected staphylococcus sepsis in settings in which MRSA is highly prevalent.

2.2. *Indirectness: differences in interventions (applicability)*

As for the population, systematic reviewers will clearly specify the interventions of interest in their eligibility criteria, ensuring that only directly relevant studies will be eligible. Again, however, there may be exceptions. For example, a systematic review might have an a priori hypothesis that a surgical procedure would have different effects when undertaken by subspecialists in referral centers compared with general surgeons in the community. If they found no studies that tested the procedure in community hospitals, review authors might conclude that the effects of the procedure undertaken by community general surgeons were uncertain.

Guideline developers may often find the best evidence addressing their question in trials of related interventions. For example, a guideline addressing the value of colonoscopic screening for colon cancer will find the randomized control trials (RCTs) of fecal occult blood screening that showed a decrease in colon cancer mortality in people receiving the intervention of relevance. Whether to rate down one or two levels in this context is a matter of judgment.

There may be instances in which the intervention differs, but authors may conclude that there is no need to rate down for quality. For example, older trials show a high efficacy of intramuscular penicillin for gonococcal infection, but guidelines might reasonably recommend alternative antibiotic regimes based on current local in vitro resistance patterns, and consider the evidence as high quality.

Interventions may be delivered differently in different settings. For instance, a systematic review of music therapies for autism found that trials tested structured approaches that are used more commonly in North America than in Europe. Because the interventions differ, the results from structured approaches are more applicable to North America and the results of less structured approaches are more applicable in Europe. Issues of setting are particularly crucial for the outcome of resource use (cost). The resources required (or at least used) for a particular intervention may vary widely across settings, and the opportunity cost (what alternatives could be purchased for the same money) differs to an even greater extent.

Guideline panelists should consider rating down the quality of the evidence if the intervention cannot be implemented with the same rigor or technical sophistication in their setting as in the RCTs from which the data come. Carotid endarterectomy provides a commonly cited example of such a situation [2]. Indirectness of this sort becomes a major issue—particularly for lower-income countries—for resource-intensive interventions. We have noted this issue under "setting" for indirectness of interventions, in which we referred to how music therapy for autism may be delivered differently in one jurisdiction than another. The same is true for other complex interventions such as rehabilitation programs and public health interventions. There may be important differences in implementation across settings that can weaken inferences regarding applicability.

As with all other aspects of rating quality of evidence, there is a continuum of similarity of the intervention that will require judgment. It is rare, and usually unnecessary, for the intended populations and interventions to be identical to those in the studies, and we should only rate down if the differences are considered sufficient to make a difference in outcome likely. For example, trials of simvastatin show cardiovascular mortality reductions: suggesting night rather than morning dosing (because of greater cholesterol reduction) would not warrant rating down for differences in the intervention. A new statin with available evidence only from lipid levels might, however, require rating down quality for indirectness, and trials of a new class of cholesterol-lowering agents in which RCTs have not addressed impact on cardiovascular events would certainly require rating down for indirectness. One could conceptualize this as

rating down for either indirectness of interventions or indirectness of outcomes.

2.3. Indirectness: differences in outcome measures (surrogate outcomes)

GRADE specifies that both those conducting systematic reviews and those developing practice guidelines should begin by specifying every important outcome of interest. The available studies may have measured the impact of the intervention of interest on outcomes related to, but different from, those of primary importance to patients.

The difference between desired and measured outcomes may relate to time frame. For example, a systematic review of behavioral and cognitive-behavioral interventions for outwardly directed aggressive behavior in people with learning disabilities showed that a program of 3-week relaxation training significantly reduced disruptive behaviors at 3 months [3]. Unfortunately, no eligible trial assessed the review authors' predefined outcome of interest, the long-term impact defined as effect at 9 months or greater. The argument for rating down becomes even stronger when one considers that other types of behavioral interventions have shown an early beneficial effect that was not sustained at 6 months follow-up [3]. When there is a discrepancy between the time frame of measurement and that of interest, whether to rate down by one or two levels will depend on the magnitude of the discrepancy. In this case, one could argue for either option.

Another source of indirectness related to measurement of outcomes is the use of substitute or surrogate endpoints in place of the patient-important outcome of interest. Table 2 lists a number of such surrogate measures that are common in current clinical investigation.

Table 3 presents the logic of patient-important and surrogate outcomes as applied to disturbances in calcium and phosphate metabolism in patients with end-stage renal disease. Hyperphosphatemia is associated with abnormal bone fragility and consequent fractures; soft tissue calcification and associated pain; coronary calcification and associated myocardial infarction; and possible increased mortality. These adverse outcomes are the important endpoints in treating the calcium/phosphate abnormalities.

Up to now, however, RCTs of alternative therapeutic interventions have focused on measures of calcium/phosphate metabolism. In general, the use of a surrogate outcome requires rating down the quality of evidence by one, or even two, levels. Consideration of the biology, mechanism, and natural history of the disease can be helpful in making a decision about indirectness. For instance, because concentrations of calcium and phosphate are far removed in the putative causal pathway from the patient-important endpoints, we would rate down the quality of evidence with respect to this outcome by two levels (Table 3). Surrogates that are closer in the putative causal pathway to the adverse outcomes are coronary calcification (for myocardial infarction), bone density (for fractures), and soft-tissue calcification (for pain), and these outcomes warrant rating down by only one level for indirectness.

A systematic review suggesting a benefit of low molecular weight heparin vs. unfractionated heparin for perioperative thromboprophylaxis in patients with cancer provides an example in which rating down by just one level for indirectness is probably appropriate. The confidence intervals (CIs) around reduction in the important outcome, symptomatic deep venous thrombosis (DVT), were extremely wide (RR 0.73; 95% CI: 0.23, 2.28). When the outcome included the surrogate, asymptomatic DVT (which provided most events), the difference in favor of low molecular weight heparin was much more convincing (RR = 0.72; 95% CI: 0.55, 0.94) [4]. Convincing evidence of reduction in asymptomatic events provides, in our view, moderate quality evidence of a reduction in symptomatic events.

Rarely, surrogates are sufficiently well established that review authors or guideline panelists should choose not to rate down quality of evidence for indirectness. In our view, this should be restricted to situations in which, within the same class of drug (e.g., beta-blockers, calcium antagonists, diuretics, bisphosphonates), changes in the surrogate have repeatedly proved closely related to changes in the patient-important outcome in the context of RCTs. One might use this rationale, for example, to justify not rating

Table 2. Examples of surrogate outcomes

Condition	Patient-important outcome(s)	Surrogate outcome(s)
Diabetes mellitus	Diabetic symptoms, hospital admission, complications (cardiovascular, eye, renal, neuropathic)	Blood glucose, A1C
Hypertension	Cardiovascular death, myocardial infarction, stroke	Blood pressure
Dementia	Patient function, behavior, caregiver burden	Cognitive function
Osteoporosis	Fractures	Bone density
Adult Respiratory Distress Syndrome	Mortality	Oxygenation
End-stage renal disease	Quality of life, morbidity (such as shunt thrombosis or heart failure), mortality	Hemoglobin
Venous thrombosis	Symptomatic venous thrombosis	Asymptomatic venous thrombosis
Chronic respiratory disease	Quality of life, exacerbations, mortality	Pulmonary function, exercise capacity
Cardiovascular disease/risk	Vascular events, mortality	Serum lipids

Table 3. Surrogate and patient-important outcomes for phosphate lowering drugs in patients with renal failure and hyperphosphatemia

	Surrogate outcomes	
Patient-important outcomes	***Indirect*** **(Lower the quality of evidence by one level)**	***Very indirect*** **(Lower the quality of evidence by two levels)**
Myocardial infarction	Coronary calcification	Measures of calcium/phosphate metabolism
Fractures	Bone density	
Pain because of soft-tissue calcification	Soft-tissue calcification	

down low-density lipoprotein (LDL) as a surrogate for coronary events in evaluating the evidence from RCTs of a new statin. One would, however, rate down for indirect outcomes the evidence from RCTs of another class of cholesterol-lowering agents (e.g., ezetimibe) if the outcome measure was LDL rather than coronary events. Even this highly restricted criterion for not rating down a surrogate (multiple randomized trials within a single drug class show a clear and consistent relationship between change in the surrogate and an effect measure such as RR reduction) may be problematic (Box 2).

Investigators may use sophisticated statistical approaches to examine the relationship between a surrogate and a patient-important outcome. For instance, investigators examined the "validity" of progression-free survival as a surrogate for overall survival for anthracycline- and taxine-based chemotherapy for advanced breast cancer [5]. They found a statistically significant association between progression-free and overall survival in the randomized trials they analyzed, but predicting overall survival using progression-free survival remained fraught with uncertainty. Rating down quality by one level for the surrogate would be appropriate in this situation.

Several groups have developed systems for rating the "validity" of a surrogate [6,7,16]. Each of these systems finds evidence from surrogates convincing only when the association has been strongly and repeatedly established in RCTs. Systematic review authors and guideline developers may wish to refer to these systems when pondering whether to rate down for indirectness of outcomes.

Box 2 Arguments against ever-considering evidence from surrogates of high quality

One might well be tempted to assume a new statin that improves lipid profiles in the same way as older statins would result in similar improvement in cardiovascular risk. Authorities have, however, raised arguments about assuming that low-density lipoprotein (LDL) reductions with a new statin will translate into the expected reduction in cardiovascular risk [13,14]. Indeed, in one large trial in hemodialysis patients, large reductions in LDL failed to effect reductions in cardiovascular events [15]. In addition, deciding what constitutes a class of drugs (e.g., all beta-blockers; all cardioselective beta-lockers; all beta-blockers with or without alpha-blocking properties) is not straightforward [16,17]. Finally, from a clinical point of view, even if one accepts that a surrogate provides high-quality evidence regarding benefit, a new agent may have a different—and highly problematic—side effect profile. Note, for instance, cerivastatin's greatly increased—relative to other statins—propensity to cause life-threatening rhabdomyolysis.

2.4. *Indirectness: indirect comparisons*

The final type of indirectness occurs when we have no direct (i.e., head-to-head) comparisons between two or more interventions of interest. For instance, consider a comparison of two active drugs, A and B. Although RCTs comparing A and B may be unavailable, RCTs may have compared A to placebo and B to placebo. Such trials allow indirect comparisons of the magnitude of effect of A and B. Such evidence is of lower quality than head-to-head comparisons of A and B would provide.

Indirect comparisons of prophylactic treatments for osteoporotic fractures illustrate the challenges of indirect comparisons. Trials of different agents suggest apparent differences in RR reduction, tempting one to attribute these differences to varying effectiveness of the drugs under consideration. The trials, however, enrolled different groups of patients; some may be more responsive than others. In addition, trials varied in criteria for diagnosis of both vertebral and nonvertebral fractures. It may be these differences, rather than differences in the effectiveness of the interventions, that are responsible for variation in RR [8]. A systematic review of different doses of aspirin illustrates the difficulties of inferences from indirect comparisons (Box 3).

The validity of the indirect comparison rests on the assumption that factors in the design of the trial (the patients, co-interventions, measurement of outcomes) and the methodological quality are not sufficiently different to result in different effects (in other words, true differences in effect explain all apparent differences). Some authors refer to this as the "similarity assumption" [9]. Because this assumption is always in some doubt, indirect comparisons always warrant rating down by one level in quality of evidence. Whether to rate down two levels depends on the plausibility that alternative factors (population, interventions, co-interventions, outcomes, and study methods) explain or obscure differences in effect. Of the many

Box 3 Difficulties making inferences from indirect comparisons: low- vs. medium-dose aspirin

A systematic review considered the relative merits of low dose (50–150 mg daily) vs. medium dose (300–325 mg daily) of aspirin to prevent graft occlusion after coronary artery bypass surgery [18]. Authors found five relevant trials that compared aspirin with placebo, of which two tested medium dose and three low-dose aspirin. The pooled relative risk (RR) of the likelihood of a graft occlusion was 0.74 (95% confidence interval [CI]: 0.60, 0.91) in the low-dose trial and 0.55 (95% CI: 0.28, 0.82) in the medium-dose trials. The RR of medium vs. low dose was 0.74 (95% CI: 0.52, 1.06; $P = 0.10$) suggesting (but not very convincingly) the possibility of a larger effect with the medium-dose regimens.

This comparison is weaker than if the randomized trials had compared the two aspirin dose regimens directly because there are other study characteristics that might be responsible for any differences found (or resulted in undetected differences that in fact exist). Compared with the low dose vs. placebo trials, in medium dose vs. placebo trials, the patients studied may be different, effective or harmful interventions other than the therapy under investigation may have been differently administered, and outcomes may have been measured differently (e.g., dissimilar criteria for events or varying duration of follow-up). Differences in study methods may also explain the results: trials with a higher risk of bias may result in smaller—or more likely larger—treatment effects.

challenging judgments that rating quality of evidence demands, this is one of the most difficult.

The judgment is made more difficult yet by the necessity to consider the statistical approaches that investigators have taken in making indirect comparisons. Simply using the results from the active arms in two or more studies is naïve and potentially misleading. More sophisticated statistical approaches that consider differences between active and control arms are more appropriate [10,11].

The comparison of low- vs. medium-dose aspirin regimens (Box 3) used a valid statistical approach to compare the RRs in one set of trials to the RR in the other set. The review authors present data suggesting that the trials enrolled patients who were very similar with respect to mean age (56–60 years), sex distribution (83–100% men), proportion of smokers (65–68% in the two trials reporting), proportion of hypertensive patients (31–53% in the four trials reporting), and mean cholesterol (5.7–7.2 mmol/L). The authors did not mention whether the two sets of studies differed in the use of a cointervention—aggressiveness of antihypertensive treatment or the use of lipid lowering agents, for instance. In terms of methods, one trial in each set standardized surgical procedures, all were blinded and included a placebo arm, two medium-dose and one low-dose trial reported formal randomization by research-coordinating centers or pharmacy, and one trial in each group reported independent angiographic assessment of vein graft patency. Both sets of trials had very high loss to follow-up (i.e., no angiography)—three of five trials reported rates of more than 50%.

On balance, we would rate down the quality of the evidence only one level for indirectness. The decision in this case has little effect on clinical decision making in that other considerations (risk of bias—high loss to follow-up, imprecision—wide CIs around the RR in moderate vs. low-dose trials, and indirectness of outcomes—graft occlusion is a surrogate for events such as myocardial infarction and cardiovascular deaths) already place this as low-quality evidence. The indirect comparison leaves us with very low-quality evidence.

Increasingly, recommendations must simultaneously address more than two interventions. For instance, possible approaches to thrombolysis in myocardial infarction include streptokinase, alteplase, reteplase, and tenecteplase [12]. Attempts to address such issues of the relative effectiveness of multiple interventions inevitably involve indirect comparisons. These meta-analyses have received different labels; currently popular terms include "network meta-analyses," "mixed treatment comparison," and "multiple treatments meta-analysis."

There are both simple, inappropriate approaches, and a number of sophisticated appropriate statistical approaches available for assessing simultaneous multiple comparisons. A variety of recently developed Bayesian statistical methods may help in generating estimates of the relative effectiveness of multiple interventions, but these methods may give different estimates. This raises the possibility of bias, and the issue of the best-quality indirect analysis is unsettled. Their confident application requires, in addition to indirect comparison evidence, substantial evidence from direct comparisons—evidence that is often unavailable [12]. Ascertaining the extent to which patients, co-interventions, measurement of outcomes, and risk of bias in studies of multiple interventions are similar presents another major challenge. Interpretation when direct and indirect evidence is inconsistent is uncertain, and may warrant rating down the direct evidence for inconsistency. A recent simultaneous treatment comparison illustrates the challenges of evaluating such studies (Box 4). The methods to conduct and assess such network meta-analyses, including GRADE's approach, remain in evolution. The coming years should see refinement in criteria for judging the quality of evidence from network meta-analyses.

A final point is that it is possible, at least in theory, for indirect comparisons to yield more accurate results than direct comparisons. This could be true if direct comparisons suffer from risk of bias that indirect comparisons do not. This may occur if the direct comparisons are conducted by those with an investment in the result (e.g., industry).

Box 4 An example of the challenges of network meta-analysis

Investigators conducted a simultaneous treatment comparison of 12 new generation antidepressants [19]. The authors evaluated 117 randomized trials involving over 25,000 patients; their article provides no information about the similarity of the patients (other than that they all had major unipolar depression), or about cointervention (behavioral therapies, for instance). In correspondence with the authors, however, they indicated that they excluded trials with treatment-resistant depression, argued that different types of depression have similar treatment responses, and that it is very likely that patients did not receive important cointervention. With respect to risk of bias, the authors tell us, using the Cochrane collaboration approach to assessing risk of bias [20] that risk of bias in most studies was "unclear," and 12 were at low risk of bias; presumably a small number was at high risk of bias. This is helpful, although "unclear" represents a wide range of risk of bias.

All studies involved head-to-head comparisons between at least two of the 12 drugs; the 117 trials involved 70 individual comparisons (e.g., two comparisons between fluoxetine and fluvoxamine). The authors reported statistically significant differences between direct and indirect comparisons in only three of 70 comparisons of drug response. The power of such tests was, however, not likely high. Overall, we would be inclined to take a cautious approach to this network meta-analysis and rate down two levels for indirectness.

3. Mechanism

Another type of indirect evidence that we have not addressed relates to mechanism of action. The GRADE system does not rate evidence either up or down based on the mechanism or pathophysiological basis of a treatment. RCTs typically begin with a reasonable expectation of success based, to some degree, on biological rationale. But judgments of exactly how strong is the rationale are easily open to dispute, and GRADE does not suggest using them directly as a basis for rating evidence quality up or down.

Mechanism does, however, have multiple roles in the evaluation of evidence: in selecting studies for systematic reviews, in the applicability of evidence to different interventions or populations, in judging whether to believe subgroup analyses, and in deciding the extent to which one rates down quality of evidence based on surrogate outcomes. Although it would make little sense to pool studies based on similar costs or color of tablets, treatments with similar mechanism are commonly meta-analyzed. Because no two studies will have exactly the same eligibility criteria and interventions, judgments based on our biological understanding are necessary to determine which studies to include in generating a single pooled estimate of effect.

Similarly, we need to make judgments based on mechanism to apply evidence about treatments. For example, if a trial that included patients aged 50–70 years showed effect, then we would undoubtedly be happy to apply the results to 49- or 71-year olds (and likely well younger than 49 and well older than 71 years) but not to children. If a study showed 5 days of antibiotics were effective, then we might be happy to use 7 days but not 3 days.

Judgments regarding surrogate outcomes may, however, be more complex. For example, consider a three-dose vaccine that reduced the incidence of the target illness. We might be happy to consider an accelerated delivery of three doses of exactly the same vaccine to be as effective as the original if the studies showed that the standard and accelerated three-dose regimes had a similar serological response (i.e., we might not rate down quality of evidence because of the surrogate outcome of serological response). However, we might rate down for use of a surrogate outcome if a new class of antihypertensive agents (e.g., the direct renin inhibitor aliskiren, recently licensed in the United States) showed a similar reduction in blood pressure to existing agents but without evidence of benefit on patient-important outcomes.

4. Simultaneous consideration of all types of indirectness

Guideline developers will usually need to consider the combined effect of all the four types of indirectness—and problems in more than one may suggest the need to rate down two levels in the quality of evidence. This consideration is not a simple additive process, but rather a judgment about whether any, and how much, rating down is warranted. In general, evidence based on surrogate outcomes should usually trigger rating down, whereas the other types of indirectness will require a more considered judgment.

References

[1] Schunemann HJ, Hill SR, Kakad M, Bellamy R, Uyeki TM, Hayden FG, et al. WHO Rapid Advice Guidelines for pharmacological management of sporadic human infection with avian influenza A (H5N1) virus. Lancet Infect Dis 2007;7(1):21–31.

[2] Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet 2005;365(9453):82–93.

[3] Hassiotis A, Hall I. Behavioural and cognitive-behavioural interventions for outwardly-directed aggressive behaviour in people with learning disabilities. Cochrane Database Syst Rev 2004;1: CD003406. DOI:10.1002/14651858.CD003406.pub2.

[4] Akl EA, Terrenato I, Barba M, Sperati F, Sempos EV, Muti P, et al. Low-molecular-weight heparin vs unfractionated heparin for

perioperative thromboprophylaxis in patients with cancer: a systematic review and meta-analysis. Arch Intern Med 2008;168:1261–9.
[5] Miksad RA, Zietemann V, Gothe R, Schwarzer R, Conrads-Frank A, Schnell-Inderst P, et al. Progression-free survival as a surrogate endpoint in advanced breast cancer. Int J Technol Assess Health Care 2008;24(4):371–83.
[6] Lassere MN, Johnson KR, Boers M, Tugwell P, Brooks P, Simon L, et al. Definitions and validation criteria for biomarkers and surrogate endpoints: development and testing of a quantitative hierarchical levels of evidence schema. J Rheumatol 2007;34:607–15.
[7] Australian Government Department of Health and Ageing. Report of the Surrogate to Final Outcome Working Group to the Pharmaceutical Benefits Advisory Committee: a framework for evaluating proposed surrogate measures and their use in submissions to PBAC. 2009.
[8] Sebba A. Comparing non-vertebral fracture risk reduction with osteoporosis therapies: looking beneath the surface. Osteoporos Int 2009;20:675–86.
[9] Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG, et al. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ 2009;338:b1147.
[10] Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol 1997;50:683–91.
[11] Glenny AM, Altman DG, Song F, Sakarovitch C, Deeks JJ, D'Amico R, et al. Indirect comparisons of competing interventions. Health Technol Assess 2005;9(26):1–134, iii–iv.
[12] Caldwell DM, Ades AE, Higgins JP. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005;331(7521):897–900.
[13] de Lorenzo F, Feher M, Martin J, Collot-Teixeira S, Dotsenko O, McGregor JL. Statin therapy-evidence beyond lipid lowering contributing to plaque stability. Curr Med Chem 2006;13:3385–93.
[14] Fisman EZ, Adler Y, Tenenbaum A. Statins research unfinished saga: desirability versus feasibility. Cardiovasc Diabetol 2005;4(1):8.
[15] Wanner C, Krane V, März W, Olschewski M, Mann JF, Ruf G, et al. Atorvastatin in patients with type 2 diabetes mellitus undergoing hemodialysis. N Engl J Med 2005;353(3):238–48.
[16] Bucher H, Kunz R, Cook D, Holbrook A, Guyatt G. Surrogate outcomes. 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.
[17] Kunz R, Bucher H, McAlister F, Holbrook A, Guyatt G. Drug class effects. 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.
[18] Lim E, Ali Z, Ali A, Routledge T, Edmonds L, Altman DG, et al. Indirect comparison meta-analysis of aspirin therapy after coronary surgery. BMJ 2003;327(7427):1309.
[19] Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet 2009;373(9665):746–58.
[20] 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.