Users' Guides to the Medical Literature

# How to Read a Systematic Review and Meta-analysis and Apply the Results to Patient Care

## Users' Guides to the Medical Literature

Mohammad Hassan Murad, MD, MPH; Victor M. Montori, MD, MSc; John P. A. Ioannidis, MD, DSc; Roman Jaeschke, MD, MSc; P. J. Devereaux, MD, PhD; Kameshwar Prasad, MD, DM, FRCPE; Ignacio Neumann, MD, MSc; Alonso Carrasco-Labra, DDS, MSc; Thomas Agoritsas, MD; Rose Hatala, MD, MSc; Maureen O. Meade, MD; Peter Wyer, MD; Deborah J. Cook, MD, MSc; Gordon Guyatt, MD, MSc

Clinical decisions should be based on the totality of the best evidence and not the results of individual studies. When clinicians apply the results of a systematic review or meta-analysis to patient care, they should start by evaluating the credibility of the methods of the systematic review, ie, the extent to which these methods have likely protected against misleading results. Credibility depends on whether the review addressed a sensible clinical question; included an exhaustive literature search; demonstrated reproducibility of the selection and assessment of studies; and presented results in a useful manner. For reviews that are sufficiently credible, clinicians must decide on the degree of confidence in the estimates that the evidence warrants (quality of evidence). Confidence depends on the risk of bias in the body of evidence; the precision and consistency of the results; whether the results directly apply to the patient of interest; and the likelihood of reporting bias. Shared decision making requires understanding of the estimates of magnitude of beneficial and harmful effects, and confidence in those estimates.

*JAMA*. 2014;312(2):171-179. doi:10.1001/jama.2014.5559

**Supplemental content** at jama.com

**CME Quiz** at jamanetworkcme.com and **CME Questions** page 186

**Author Affiliations:** Author affiliations are listed at the end of this article.

**Corresponding Author:** M. Hassan Murad, MD, MPH, Mayo Clinic, 200 1st St, SW, Rochester, MN 55905 (murad.mohammad@mayo.edu).

## Clinical Scenario

You are consulted regarding the perioperative management of a 66-year-old man undergoing hip replacement. He is a smoker and has a history of type 2 diabetes and hypertension. Because he has multiple cardiovascular risk factors, you consider using perioperative β-blockers to reduce the risk of postoperative cardiovascular complications. You identify a recently published systematic review and meta-analysis evaluating the effect of perioperative β-blockers on death, nonfatal myocardial infarction, and stroke.[1] How should you use this meta-analysis to help guide your clinical decision making?

## Introduction and Definitions

Traditional, unstructured review articles are useful for obtaining a broad overview of a clinical condition but may not provide a reliable and unbiased answer to a focused clinical question. A systematic review is a research summary that addresses a focused clinical question in a structured, reproducible manner. It is often, but not always, accompanied by a meta-analysis, which is a statistical pooling or aggregation of results from different studies providing a single estimate of effect. **Box 1** summarizes the typical process of a systematic review and meta-analysis including the safeguards against misleading results.

In 1994, a Users' Guide on how to use an "overview article"[2] was published in *JAMA* and presented a framework for critical appraisal of systematic reviews. In retrospect, this framework did not distinguish between 2 very different issues: the rigor of the review methods and the confidence in estimates (quality of evidence) that the results warrant. The current Users' Guide reflects the evolution of thinking since that time and presents a contemporary conceptualization.

We refer to the first judgment as the credibility[3] of the review: the extent to which its design and conduct are likely to have protected against misleading results.[4] Credibility may be undermined by inappropriate eligibility criteria, inadequate literature search, or failure to optimally summarize results. A review with credible methods, however, may leave clinicians with low confidence in effect estimates. Therefore, the second judgment addresses the confidence in estimates.[5] Common reasons for lower confidence include high risk of bias of the individual studies; inconsistent results; and small sample size of the body of evidence, leading to imprecise estimates. This Users' Guide presents criteria for judging both credibility and confidence in the estimates (**Box 2**).

This guide focuses on a question of therapy and is intended for clinicians applying the results to patient care. It does not provide comprehensive advice to researchers on how to conduct[6] or report[7] reviews. We also provide a rationale for seeking systematic reviews and meta-analyses and explaining the summary estimate of a meta-analysis.

**Box 1. The Process of Conducting a Systematic Review and Meta-analysis**

1. Formulate the question
2. Define the eligibility criteria for studies to be included in terms of Patient, Intervention, Comparison, Outcome (PICO), and study design
3. Develop a priori hypotheses to explain heterogeneity
4. Conduct search
5. Screen titles and abstracts for inclusion
6. Review full text of possibly eligible studies
7. Assess the risk of bias
8. Abstract data
9. When meta-analysis is performed:
   - Generate summary estimates and confidence intervals
   - Look for explanations of heterogeneity
   - Rate confidence in estimates of effect

## Why Seek Systematic Reviews and Meta-analysis?

When searching for evidence to answer a clinical question, it is preferable to seek a systematic review, especially one that includes a meta-analysis. Single studies are liable to be unrepresentative of the total evidence and be misleading.[8] Collecting and appraising multiple studies require time and expertise that practitioners may not have. Systematic reviews include a greater range of patients than any single study, potentially enhancing confidence in applying the results to the patient at hand.

Meta-analysis of a body of evidence includes a larger sample size and more events than any individual study, leading to greater precision of estimates, facilitating confident decision making. Meta-analysis also provides an opportunity to explore reasons for inconsistency among studies.

A key limitation of systematic reviews and meta-analyses is that they produce estimates that are as reliable as the studies summarized. A pooled estimate derived from meta-analysis of randomized trials at low risk of bias will always be more reliable than that derived from a meta-analysis of observational studies or of randomized trials with less protection against bias.

## First Judgment: Was the Methodology of the Systematic Review Credible?

### Did the Review Explicitly Address a Sensible Clinical Question?

Systematic reviews of therapeutic questions should have a clear focus and address questions defined by particular patients, interventions, comparisons, and outcomes (PICO). When a meta-analysis is conducted, the issue of how narrow or wide the scope of the question becomes particularly important. Consider 4 hypothetical examples of meta-analyses with varying scope: (1) the effect of all cancer treatments on mortality or disease progression; (2) the effect of chemotherapy on prostate cancer–specific mortality; (3) the effect of docetaxel in castration-resistant prostate cancer on cancer-specific mortality; (4) the effect of docetaxel in metastatic castration-resistant prostate cancer on cancer-specific mortality

These 4 questions represent a gradually narrowing focus in terms of patients, interventions, and outcomes. Clinicians will be uncomfortable with a meta-analysis of the first question and likely of the second. Combining the results of these studies would yield an estimate of effect that would make little sense or be misleading. Comfort level in combining studies increases in the third and fourth questions, although clinicians may even express concerns about the fourth question because it combines symptomatic and asymptomatic populations.

**Box 2. Guide for Appraising and Applying the Results of a Systematic Review and Meta-analysis[a]**

**First Judgment: Evaluate the Credibility of the Methods of Systematic Review**

Did the review explicitly address a sensible clinical question?

Was the search for relevant studies exhaustive?

Were selection and assessments of studies reproducible?

Did the review present results that are ready for clinical application?

Did the review address confidence in estimates of effect?

**Second Judgment: Rate the Confidence in the Effect Estimates**

How serious is the risk of bias in the body of evidence?

Are the results consistent across studies?

How precise are the results?

Do the results directly apply to my patient?

Is there concern about reporting bias?

Are there reasons to increase the confidence rating?

[a] Systematic reviews can address multiple questions. This guide is applied to aspects of the systematic review that answer the clinical question at hand—ideally the effect of the intervention vs the comparator of interest on all outcomes of importance to patients.

What makes a meta-analysis too broad or too narrow? Clinicians need to decide whether, across the range of patients, interventions or exposures, and outcomes, it is plausible that the intervention will have a similar effect. This decision will reflect an understanding of the underlying biology and may differ between individuals; it will only be possible, however, when systematic reviewers explicitly present their eligibility criteria.

### Was the Search for Relevant Studies Exhaustive?

Systematic reviews are at risk of presenting misleading results if they fail to secure a complete or representative sample of the available eligible studies. For most clinical questions, searching a single database is insufficient. Searching MEDLINE, EMBASE, and the Cochrane Central Register of Controlled Trials may be a minimal requirement for most clinical questions[6] but for many questions will not uncover all eligible articles. For instance, one study demonstrated that searching MEDLINE and EMBASE separately retrieved, respectively, only 55% and 49% of the eligible trials.[9] Another study found that 42% of published meta-analyses included at least 1 trial not indexed in MEDLINE.[10] Multiple synonyms and search terms to describe each concept are needed.

Additional references are identified through searching trial registries, bibliography of included studies, abstract presentations, contacting experts in the field, or searching databases of pharmaceutical companies and agencies such as the US Food and Drug Administration.

### Were Selection and Assessments of Studies Reproducible?

Systematic reviewers must decide which studies to include, the extent of risk of bias, and what data to abstract. Although they

follow an established protocol, some of their decisions will be subjective and prone to error. Having 2 or more reviewers participate in each decision may reduce error and subjectivity. Systematic reviewers often report a measure of agreement on study selection and quality appraisal (eg, κ statistic). If there is good agreement between the reviewers, the clinician can have more confidence in the process.

### Did the Review Present Results That Are Ready for Clinical Application?

Meta-analyses provide estimates of effect size (the magnitude of difference between groups).[11] The type of effect size depends on the nature of the outcome (relative risk, odds ratio, differences in risk, hazard ratios, weighted mean difference, and standardized mean difference). Standardized effect sizes are expressed in multiples of the standard deviation. This facilitates comparison of studies, irrespective of units of measure or the measurement scale.

Results of meta-analyses are usually depicted in a forest plot. The point estimate of each study is typically presented as a square with a size proportional to the weight of the study, and the confidence interval (CI) is presented as a horizontal line. The combined summary effect, or pooled estimate, is typically presented as a diamond, with its width representing the confidence or credible interval (the CI indicates the range in which the true effect is likely to lie). Forest plots for the perioperative β-blockers scenario are shown in the **Figure**.

Meta-analysis provides a weighted average of the results of the individual studies in which the weight of the study depends on its precision. Studies that are more precise (ie, have narrower CIs) will have greater weight and thus more influence on the combined estimate. For binary outcomes such as death, the precision depends on the number of events and sample size. In panel B of the Figure, the POISE trial[12] had the largest number of deaths (226) and the largest sample size (8351); therefore, it had the narrowest CI and the largest weight (the effect from the trial is very similar to the combined effect). Smaller trials with smaller numbers of events in that plot have a much wider CI, and their effect size is quite different from the combined effect (ie, had less weight in meta-analysis). The weighting of continuous outcomes is also based on the precision of the study, which in this case depends on the sample size and SD (variability) of each study.

In most meta-analyses such as the one in this clinical scenario, aggregate data from each study are combined (ie, study-level data). When data on every individual enrolled in each of the studies are available, individual-patient data meta-analysis is conducted. This approach facilitates more detailed analysis that can address issues such as true intention-to-treat and subgroup analyses.

Relative association measures and continuous outcomes pose challenges to risk communication and trading off benefits and harms. Patients at high baseline risk can expect more benefit than those at lower baseline risk from the same intervention (the same relative effect). Meta-analysis authors can facilitate decision making by providing absolute effects in populations with various risk levels.[13,14] For example, given 2 individuals, one with low Framingham risk of cardiovascular events (2%) and the other with a high risk (28%), we can multiply each of these baseline risks with the 25% relative risk reduction obtained from a meta-analysis of statin therapy trials.[15] The resulting absolute risk reduction (ie, risk difference) attributable to statin therapy would be 0.5% for the low-risk individual and 7% for the high-risk individual.

Continuous outcomes can also be presented in more useful ways. Improvement of a dyspnea score by 1.06 scale points can be better understood by informing readers that the minimal amount considered by patients to be important on that scale is 0.5 points.[16] A standardized effect size (eg, paroxetine reduced depression severity by 0.31 SD units) can be better understood if (1) referenced to cutoffs of 0.2, 0.5, and 0.8 that represent small, moderate, and large effect, respectively; (2) translated back to natural units with which clinicians have more familiarity (eg, converted to a change of 2.47 on the Hamilton Rating Scale for Depression); or (3) dichotomized (for every 100 patients treated with paroxetine, 11 will achieve important improvement).[17]

### Did the Review Address Confidence in Estimates of Effect?

A well-conducted (ie, credible) systematic review should present readers with information needed to make their second judgement: the confidence in the effect estimates. For example, if systematic reviewers do not evaluate the risk of bias in the individual studies or attempt to explain heterogeneity, this second judgement will not be possible.

In **Box 3**, we return to the clinical scenario to determine credibility of the systematic review identified. Overall, you conclude that the credibility of the methods of this systematic review is high and move on to examine the estimates of effect and the associated confidence in these estimates.

## Second Judgment: What Is the Confidence in the Estimates of Effect?

Several systems are used to evaluate the quality of evidence, of which 4 are most commonly used: the Grading of Recommendations Assessment, Development and Evaluation (GRADE) and the systems from the American Heart Association, the US Preventive Services Task Force, and the Oxford Centre for Evidence-Based Medicine.[5,18-20] These systems share the similar features of being used by multiple organizations and providing a confidence rating in the estimates that gives randomized trials a higher rating than nonrandomized studies. The 4 systems are described in eTable 1 in the Supplement.

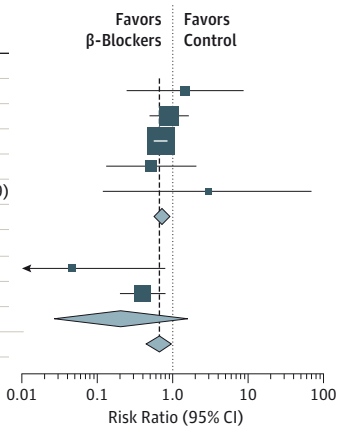
The general framework used in this Users' Guide follows the GRADE approach.[21] GRADE categorizes confidence in 4 categories: high, moderate, low, or very low. The lower the confidence, the more likely the underlying true effect is substantially different from the observed estimate of effect and, thus, the more likely that further research would demonstrate different estimates.[5]

Confidence ratings begin by considering study design. Randomized trials are initially assigned high confidence and observational studies are given low confidence, but a number of factors may modify these initial ratings. Confidence may decrease when there is high risk of bias, inconsistency, imprecision, indirectness, or concern about publication bias. An increase in confidence rating is uncommon and occurs primarily in observational studies when the effect size is large. Readers of a systematic review can consider these factors regardless of whether systematic review authors formally used this approach. Readers do, however, require the necessary information, and thus the need for a final credibility guide: Did the Review Address Confidence in Estimates of Effect?

Figure. Results of a Meta-analysis of the Outcomes of Nonfatal Infarction, Death, and Nonfatal Stroke in Patients Receiving Perioperative β-Blockers

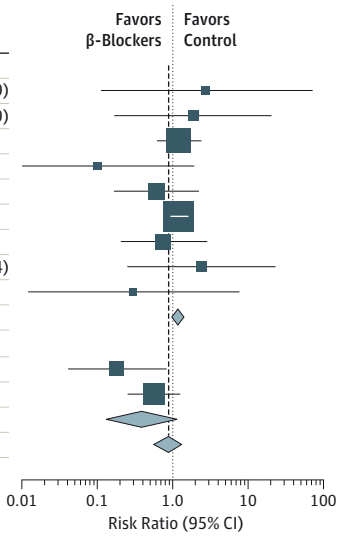
A Nonfatal myocardial infarction

| Source | β-Blockers Events, No. | β-Blockers Total, No. | Control Events, No. | Control Total, No. | RR (95% CI) |
|---|---|---|---|---|---|
| Low risk of bias | | | | | |
| DIPOM | 3 | 462 | 2 | 459 | 1.49 (0.25-8.88) |
| MaVS | 19 | 246 | 21 | 250 | 0.92 (0.51-1.67) |
| POISE | 152 | 4174 | 215 | 4177 | 0.71 (0.58-0.87) |
| POBBLE | 3 | 55 | 5 | 48 | 0.52 (0.13-2.08) |
| BBSA | 1 | 110 | 0 | 109 | 2.97 (0.12-72.19) |
| Subtotal ($I^2$ = 0%; $P$ = .70) | | | | | 0.73 (0.61-0.88) |
| High risk of bias | | | | | |
| Poldermans | 0 | 59 | 9 | 53 | 0.05 (0.00-0.79) |
| Dunkelgrun | 11 | 533 | 27 | 533 | 0.41 (0.20-0.81) |
| Subtotal ($I^2$ = 57%; $P$ = .13) | | | | | 0.21 (0.03-1.61) |
| Overall | | | | | 0.67 (0.47-0.96) |
| $I^2$ = 29%; $P$ = .21 | | | | | |
| Interaction test between groups, $P$ = .22 | | | | | |



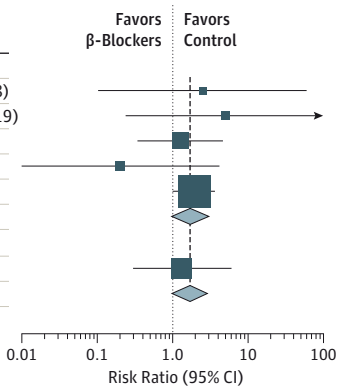

B Death

| Source | β-Blockers Events, No. | β-Blockers Total, No. | Control Events, No. | Control Total, No. | RR (95% CI) |
|---|---|---|---|---|---|
| Low risk of bias | | | | | |
| BBSA | 1 | 110 | 0 | 109 | 2.97 (0.12-72.19) |
| Bayliff | 2 | 49 | 1 | 50 | 2.04 (0.19-21.79) |
| DIPOM | 20 | 462 | 15 | 459 | 1.32 (0.69-2.55) |
| MaVS | 0 | 246 | 4 | 250 | 0.11 (0.01-2.09) |
| Neary | 3 | 18 | 5 | 20 | 0.67 (0.19-2.40) |
| POISE | 129 | 4174 | 97 | 4177 | 1.33 (1.03-1.73) |
| Mangano | 4 | 99 | 5 | 101 | 0.82 (0.23-2.95) |
| POBBLE | 3 | 55 | 1 | 48 | 2.62 (0.28-24.34) |
| Yang | 0 | 51 | 1 | 51 | 0.33 (0.01-8.00) |
| Subtotal ($I^2$ = 0%; $P$ = .68) | | | | | 1.27 (1.01-1.60) |
| High risk of bias | | | | | |
| Poldermans | 2 | 59 | 9 | 53 | 0.20 (0.05-0.88) |
| Dunkelgrun | 10 | 533 | 16 | 533 | 0.63 (0.29-1.36) |
| Subtotal ($I^2$ = 44%; $P$ = .18) | | | | | 0.42 (0.15-1.23) |
| Overall | | | | | 0.94 (0.63-1.40) |
| $I^2$ = 30%; $P$ = .16 | | | | | |
| Interaction test between groups, $P$ = .04 | | | | | |




C Nonfatal stroke

| Source | β-Blockers Events, No. | β-Blockers Total, No. | Control Events, No. | Control Total, No. | RR (95% CI) |
|---|---|---|---|---|---|
| Low risk of bias | | | | | |
| POBBLE | 1 | 53 | 0 | 44 | 2.50 (0.10-59.88) |
| DIPOM | 2 | 462 | 0 | 459 | 4.97 (0.24-103.19) |
| MaVS | 5 | 246 | 4 | 250 | 1.27 (0.35-4.67) |
| Yang | 0 | 51 | 2 | 51 | 0.20 (0.01-4.07) |
| POISE | 27 | 4174 | 14 | 4177 | 1.93 (1.01-3.68) |
| Subtotal ($I^2$ = 0%; $P$ = .60) | | | | | 1.73 (1.00-2.99) |
| High risk of bias | | | | | |
| Dunkelgrun | 4 | 533 | 3 | 533 | 1.33 (0.30-5.93) |
| Overall | | | | | 1.67 (1.00-2.80) |
| $I^2$ = 0%; $P$ = .71 | | | | | |
| Interaction test between groups, $P$ = .75 | | | | | |




Abbreviations: BBSA, Beta Blocker in Spinal Anesthesia study; DIPOM, Diabetic Postoperative Mortality and Morbidity trial; MaVS, Metoprolol after Vascular Surgery study; POBBLE, Perioperative β-blockade trial; POISE, Perioperative Ischemic Evaluation trial. Dotted line indicates no effect. Dashed line is centered on meta-analysis pooled estimate.

### How Serious Is the Risk of Bias in the Body of Evidence?

A well-conducted systematic review should always provide readers with insight about the risk of bias in each individual study and overall.[6,7] Differences in studies' risk of bias can explain important differences in results.[22] Less rigorous studies sometimes overestimate the effectiveness of therapeutic and preventive interventions.[23] The effects of antioxidants on the risk of prostate cancer[24] and on atherosclerotic plaque formation[25] are 2 of many

**Box 3. Using the Guide: Judgment 1, Determining Credibility of the Methods of a Systematic Review (Perioperative β-Blockers in Noncardiac Surgery)[1]**

Systematic review authors constructed a sensibly structured clinical question (in patients at higher-than-average cardiovascular risk undergoing noncardiac surgery, what is the effect of β-blockers vs no β-blockers on nonfatal myocardial infarction, death, and stroke)

They conducted a comprehensive search of numerous databases and registries

Two independent reviewers selected eligible trials, although the authors did not report extent of agreement

The authors ultimately presented results in a transparent and understandable way. Although they did not report an absolute effect—an important limitation—the raw data allow readers to easily calculate an absolute effect and a number needed to treat (Box 4 and Table).

The authors provided the information needed to address confidence in study results. They described the risk of bias for each trial, noted substantial heterogeneity in estimates of the effect of β-blockers on death, determined that risk of bias provided a likely explanation for the variability, and therefore focused on the results of the studies with low risk of bias.

examples of observational studies that showed misleading results subsequently contradicted by large randomized clinical trials.

Ideally, systematic reviewers will evaluate and report the risk of bias for each of the important outcomes measured in each individual study. There is no one correct way to assess the risk of bias.[26] Review authors can use detailed checklists or focus on a few key aspects of the study. Different study designs require the use of different instruments (eg, for randomized clinical trials, the Cochrane Risk of Bias Tool[27]). A judgment about the overall risk of bias for all of the included studies may then result in decreasing the confidence in estimates.[5]

### Are the Results Consistent Across Studies?

Readers of a meta-analysis that combines results from multiple studies should judge the extent to which results differ from study to study (ie, variability or heterogeneity). They can start by visually inspecting a forest plot,[28] first noting differences in the point estimates and then the extent to which CIs overlap. Large differences in point estimates or CIs that do not overlap suggest that random error is an unlikely explanation of the different results and therefore decreases confidence in the combined estimate.

Authors of a meta-analysis can help readers by conducting statistical evaluation of heterogeneity (eTable 2 in the Supplement). The first test is called the Cochran Q test (a yes-or-no test), in which the null hypothesis is that the underlying effect is the same in each of the studies[29] (eg, the relative risk derived from study 1 is the same as that from studies 2, 3, and 4). A low *P* value of the test means that random error is an unlikely explanation for the differences in results from study to study, thus decreasing confidence in a single summary estimate.

The $I^2$ statistic focuses on the magnitude of variability rather than its statistical significance.[30] An $I^2$ of 0% suggests that chance explains variability in the point estimates, and clinicians can be comfortable with a single summary estimate. As the $I^2$ increases, we become progressively less comfortable with unexplained variability in results.

When substantial heterogeneity exists, clinicians should look for possible explanations. Authors of meta-analyses may conduct subgroup analyses to explain heterogeneity. Such analyses may not reflect true subgroup differences, and a Users' Guide is available to aid readers in evaluating the credibility of these analyses.[7] Authors of meta-analyses can address one important credibility criterion, whether chance can explain differences between subgroups, using what is called a test of interaction.[31] The lower the *P* value of the test of interaction, the less likely chance explains the difference between intervention effects in the subgroups examined, and therefore the greater likelihood that the subgroup effect is real.

Another approach to exploring causes of heterogeneity in meta-analysis is meta-regression. Investigators construct a regression model in which independent variables are individual study characteristics (eg, the population, how the intervention was administered) and the dependent variable is the estimate of effect in each study. Conclusions from meta-regression have the same limitations as those from subgroup analysis, and inferences about explanations of heterogeneity may not be accurate. For example, meta-regression[32] of trials evaluating statin therapy in patients undergoing percutaneous interventions for acute coronary syndrome showed that the earlier statins were given, the lower the risk of cardiac events. Although the trials were randomized (to statin vs no statin or a lower-dose statin), the conclusion about early administration was not based on randomization and should be evaluated using the Users' Guide on subgroup analysis.[7]

It is not uncommon that a large degree of between-study heterogeneity remains unexplained. Clinicians and patients still need, however, a best estimate of the treatment effect to inform their decisions. Pending further research that may explain the observed heterogeneity, the summary estimate remains the best estimate of the treatment effect. Clinicians and patients must use this best available evidence, although this inconsistency between studies appreciably reduces confidence in the summary estimate.[33]

In the β-blocker meta-analysis, the risk of bias explains variability in results in the outcome of death (Figure, panel B). Results are very different for the trials with high and low risk of bias, and the *P* value for the test of interaction (.04) tells us that chance is an unlikely explanation for the difference. Therefore, we use the results from the trials with low risk of bias as our best estimate of the treatment effect.

### How Precise Are the Results?

There are 2 fundamental reasons that studies mislead: one is systematic error (otherwise known as bias), and the other is random error. Random error is large when sample sizes, and numbers of events, are small, and decreases as sample size and number of events increase. When sample size and number of events are small, we refer to results as "imprecise"; when they are large, we label results as "precise."

When results are imprecise, we lose confidence in estimates of effect. But how is the clinician to determine if results are sufficiently precise? Meta-analysis generates not only an estimate of the average effect across studies, but also a CI around that estimate. Examination of that CI—the range of values within which the true effect plausibly lies—allows a judgement of whether a meta-analysis yields results that are sufficiently precise.

Clinicians can judge precision by considering the upper and lower boundaries of the CI and then considering how they would advise

### Box 4. Using the Guide: Judgment 2, Determining the Confidence in the Estimates (Perioperative β-Blockers in Noncardiac Surgery)[1]

See the Table for the raw data used in this discussion.

#### How to Calculate Risk Difference (Absolute Risk Reduction or Increase)?

In the Figure, the risk ratio (RR) for nonfatal myocardial infarction is 0.73. The baseline risk (risk without perioperative β-blockers) can be obtained from the trial that is the largest and likely enrolled most representative population[12] (215/4177, approximately 52 per 1000). The risk with intervention would be (52/1000 × 0.73, approximately 38 per 1000). The absolute risk difference would be (52/1000 − 38/1000 = −14, approximately 14 fewer myocardial infarctions per 1000). The same process can be used to calculate the confidence intervals around the risk difference, substituting the boundaries of the confidence interval (CI) of the RR for the point estimate.

The number needed to treat to prevent 1 nonfatal myocardial infarction can also be calculated as the inverse of the absolute risk difference (1/0.014 = 72 patients).

#### Risk of Bias

Of the 11 trials included in the analysis, 2 were considered to have high risk of bias.[35,36] Limitations included lack of blinding, stopping early because of large apparent benefit,[36] and concerns about the integrity of the data.[1] The remaining 9 trials had adequate bias protection measures and represented a body of evidence that was at low risk of bias.

#### Inconsistency

Visual inspection of forest plots (Figure) shows that the point estimates, for both nonfatal myocardial infarction and death, substantially differ across studies. For the outcome of stroke, results are extremely consistent. There is minimal overlap of CIs of point estimates for the analysis of death. Confidence intervals in the analysis of nonfatal myocardial infarction do overlap to a great extent and fully overlap in the outcome of stroke. Heterogeneity *P* values were .21 for nonfatal myocardial infarction, .16 for death, and .71 for stroke; $I^2$ values were 29%, 30%, and 0%, respectively. A test of interaction between the 2 groups of studies (high risk of bias vs low risk of bias) yields a nonsignificant *P* value of .22 for myocardial infarction (suggesting that the difference between these 2 subgroups of studies could be attributable to chance) and a significant *P* value of .04 for the outcome of death. Considering that the observed heterogeneity is at least partially explained by the risk of bias and that the trials with low risk of bias for all outcomes are consistent, you decide to obtain the estimates of effect from the trials with low risk of bias and do not lower the confidence rating because of inconsistency.

#### Imprecision

For the outcomes of death and nonfatal stroke, clinical decisions would differ if the upper vs the lower boundaries of the CI represented the truth; therefore, imprecision makes us lose confidence in both estimates. No need to lower the confidence rating for nonfatal myocardial infarction.

#### Indirectness

The age of the majority of patients enrolled across the trials ranged between 50 and 70, similar to the patient in the opening scenario, who is 66 years old. Most of the trials enrolled patients with risk factors for heart disease undergoing surgical procedures classified as intermediate surgical risk, similar to the risk factors and hip surgery of the patient. Although the drug used and the dose varied across trials, the consistent results suggest we can use a modest dose of the β-blocker with which we are most familiar. The outcomes of death, nonfatal stroke, and nonfatal infarction are the key outcomes of importance to patients. Overall, the available evidence presented in the systematic review is direct and applicable to the patient of interest and addresses the key outcomes.

#### Reporting Bias

The authors of the systematic review and meta-analysis constructed funnel plots that appear to be symmetrical and results of the statistical tests for the symmetry of the plot were nonsignificant, leaving no reason for lowering the confidence rating because of possible reporting or publication bias.

#### Confidence in the Estimates

Overall, evidence warranting high confidence suggests that individuals with risk factors for heart disease can expect a reduction in risk of a perioperative nonfatal infarction of 14 in 1000 (from approximately 20 per 1000 to 6 per 1000). Unfortunately, they can also expect an increase in their risk of dying or having a nonfatal stroke. Because most people are highly averse to stroke and death, it is likely that the majority of patients faced with this evidence would decline β-blockers as part of their perioperative regimen. Indeed, that is what this patient decides when informed about the evidence.

Table. Evidence Summary of the Perioperative β-Blockers Question

| Outcome | No. of Participants (Trials) | Confidence | Relative Effect (95% CI) | Risk Difference per 1000 Patients[a] |
|---|---|---|---|---|
| Nonfatal myocardial infarction | 10 189 (5) | High | 0.73 (0.61-0.88) | 14 fewer (6 fewer to 20 fewer) |
| Stroke | 10 186 (5) | Moderate | 1.73 (1.00- 2.99) | 2 more (0 more to 6 more) |
| Death | 10 529 (9) | Moderate | 1.27 (1.01-1.60) | 6 more (0 more to 13 more) |

[a] See Box 4.

their patients were the upper boundary to represent the truth and how they would advise their patients were the lower boundary to represent the truth. If the advice would be the same in either case, then the evidence is sufficiently precise. If decisions would change across the range of the confidence interval, then confidence in the evidence will decrease.[34]

For instance, consider the results of nonfatal myocardial infarction in the β-blocker example (**Box 4** and **Table**). The CI around the absolute effect of β-blockers is a reduction of from 6 (the minimum) to 20 (the maximum) infarctions in 1000 patients given β-blockers. Considering this range of plausible effects, clinicians must ask themselves: Would my patients make different choices about

the use of β-blockers if their risk of infarction decreased by only 6 in 1000 or by as much as 20 in 1000?

One might readily point out that this judgment is subjective—it is a matter of values and preferences. Quite so, but that is the nature of clinical decision making: the trade-off between the desirable and undesirable consequences of the alternative courses of action is a matter of values and preferences and is therefore subjective. To the extent that clinicians are confident that patients would place similar weight on reductions of 6 and 20 in 1000 infarctions, concern about imprecision will be minimal. To the extent that clinicians are confident that patients will view 6 in 1000 as trivial and 20 in 1000 as important, concern about imprecision will be large. To the extent that clinicians are uncertain of their patients' values and preferences on the matter, judgments about imprecision will be similarly insecure.

The judgment regarding myocardial infarction may leave clinicians with doubt about imprecision—much less so for stroke and death (Box 4 and Table). With regard to both, if the boundary most favoring β-blockers (ie, no increase in death and stroke) represented the truth, patients would have no reluctance regarding use of β-blockers. On the other hand, if risk of death and stroke increased by, respectively, 13 and 6, reluctance regarding use of β-blockers would increase substantially. Given uncertainty about which extreme represents the truth, confidence in estimates decreases because of imprecision.

### Do the Results Directly Apply to My Patient?

The optimal evidence for decision making comes from research that directly compared the interventions in which we are interested, evaluated in the populations in which we are interested, and measured outcomes important to patients. If populations, interventions, or outcomes in studies differ from those of interest, the evidence can be viewed as indirect.

A common example of indirectness of population is when we treat a very elderly patient using evidence derived from trials that excluded elderly persons. Indirectness of outcomes occurs when trials use surrogate end points (eg, hemoglobin $A_{1c}$ level), whereas patients are most concerned about other outcomes (eg, macrovascular and microvascular disease).[37] Indirectness also occurs when clinicians must choose between interventions that have not been tested in head-to-head comparisons.[38] For instance, many trials have compared osteoporosis drugs with placebo, but very few have compared them directly against one another.[39] Making comparisons between treatments under these circumstances requires extrapolation from existing comparisons and multiple assumptions.[40]

Decisions regarding indirectness of patients and interventions depend on an understanding of whether biologic or social factors are sufficiently different that one might expect substantial differences in the magnitude of effect. Indirectness can lead to lowering confidence in the estimates.[38]

### Is There Concern About Reporting Bias?

When researchers base their decision to publish certain material on the magnitude, direction, or statistical significance of the results, a systematic error called reporting bias occurs. This is the most difficult type of bias to address in systematic reviews. When an entire study remains unreported, the standard term is publication bias. It has been shown that the magnitude and direction of results may be more important determinants of publication than study design, relevance, or quality[41] and that positive studies may be as much as 3 times more likely to be published than negative studies.[42] When authors or study sponsors selectively report specific outcomes or analyses, the term selective outcome reporting bias is used.[43]

Empirical evidence suggests that half of the analysis plans of randomized trials are different in protocols than in published reports.[44] Reporting bias can create misleading estimates of effect. A study of the US Food and Drug Administration reports showed that they often included numerous unpublished studies and that the findings of these studies can alter the estimates of effect.[45] Data on 74% of patients enrolled in the trials evaluating the antidepressant reboxetine were unpublished. Published data overestimated the benefit of reboxetine vs placebo by 115% and vs other antidepressants by 23%, and also underestimated harm.[46]

Detecting publication bias in a systematic review is difficult. When it includes a meta-analysis, a common approach is to examine whether the results of small studies differ from those of larger ones. In a figure that relates the precision (as measured by sample size, SE, or variance) of studies included in a meta-analysis to the magnitude of treatment effect, the resulting display should resemble an inverted funnel (eFigure, panel A in the Supplement). Such funnel plots should be symmetric around the combined effect. A gap or empty area in the funnel suggests that studies may have been conducted and not published (eFigure, panel B in the Supplement). Other explanations for asymmetry are, however, possible. Small studies may have a higher risk of bias explaining their larger effects, may have enrolled a more responsive patient group, or may have administered the intervention more meticulously. Last, there is always the possibility of a chance finding.

Several empirical tests have been developed to detect publication bias. Unfortunately, all have serious limitations, require a large number of studies (ideally 30 or more),[47] and none has been validated against a criterion standard of real data in which we know whether bias existed.[47]

More compelling than any of these theoretical exercises is the success of systematic reviewers in obtaining the results of unpublished studies. Prospective study registration with accessible results may be a solution to reporting bias.[48,49] Until complete reporting becomes a reality,[50] clinicians using research reports to guide their practice must remain cognizant of the dangers of reporting biases and, when they suspect bias, should lower their confidence in the estimates.[51]

### Are There Reasons to Increase the Confidence Rating?

Some uncommon situations warrant an increase in the confidence rating of effect estimates from observational studies. Consider our confidence in the effect of hip replacement on reducing pain and functional limitations in severe osteoarthritis, epinephrine to prevent mortality in anaphylaxis, insulin to prevent mortality in diabetic ketoacidosis, or dialysis to prolong life in patients with end-stage renal failure.[52] In each of these situations, we observe a large treatment effect achieved over a short period among patients with a condition that would have inevitably worsened in the absence of an intervention. This large effect can increase confidence in a true association.[52]

Box 4 and the Table summarize the effect of β-blockers in patients undergoing noncardiac surgery and addresses our confidence in the apparent effects of the intervention.

## Conclusions

Clinical and policy decisions should be based on the totality of the best evidence and not the results of individual studies. Systematic summaries of the best available evidence are required for optimal clinical decision making. Applying the results of a systematic review and meta-analysis includes a first step in which we judge the credibility of the methods of the systematic review and a second step in which we decide how much confidence we have in the estimates of effect.


### ARTICLE INFORMATION

**Author Affiliations:** Division of Preventive Medicine and Knowledge and Evaluation Research Unit, Mayo Clinic, Minnesota (Murad); Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, Minnesota (Montori); Departments of Medicine and Health Research and Policy, Stanford University School of Medicine; Department of Statistics, Stanford University School of Humanities and Sciences; and Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California (Ioannidis); Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Jaeschke); Departments of Medicine and Clinical Epidemiology and Biostatistics and Population Health Research Institute, McMaster University, Hamilton, Ontario, Canada (Devereaux); All India Institute of Medical Sciences, New Delhi (Prasad); Department of Internal Medicine, School of Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile (Neumann); Evidence-Based Dentistry Unit, Faculty of Dentistry, Universidad de Chile, Santiago, Chile, and Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Carrasco-Labra); Department Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Agoritsas); Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada (Hatala); Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Meade); Columbia University Medical Center, New York, New York (Wyer); Departments of Medicine and Clinical Epidemiology and Biostatistics and Population Health Research Institute, McMaster University, Hamilton, Ontario, Canada (Cook); Departments of Medicine and Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada (Guyatt).

**Author Contributions:** Dr Murad had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

**Conflict of Interest Disclosures:** All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Drs Murad and Montori receive funding from several non-for-profit organizations to conduct systematic reviews and meta-analysis. The Meta-Research Innovation Center at Stanford (METRICS), directed by Dr Ioannidis, is funded by a grant from the Laura and John Arnold Foundation. Dr Agoritsas was financially supported by Fellowship for Prospective Researchers Grant No. PBGEP3-142251 from the Swiss National Science Foundation. Dr Cook is a Research Chair of the Canadian Institutes of Health Research.

**Disclaimer:** Drs Murad, Montori, Jaeschke, Prasad, Carrasco-Labra, Neumann, Agoritsas, and Guyatt are members of the GRADE Working Group.


### REFERENCES


1. Bouri S, Shun-Shin MJ, Cole GD, Mayet J, Francis DP. Meta-analysis of secure randomised controlled trials of β-blockade to prevent perioperative death in non-cardiac surgery. *Heart*. 2014;100(6):456-464.

2. Oxman AD, Cook DJ, Guyatt GH; Evidence-Based Medicine Working Group. Users' guides to the medical literature, VI: how to use an overview. *JAMA*. 1994;272(17):1367-1371.

3. Alkin M. *Evaluation Roots: Tracing Theorists' Views and Influences*. Thousand Oaks, California: Sage Publications Inc; 2004.

4. Oxman AD. Checklists for review articles. *BMJ*. 1994;309(6955):648-651.

5. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines, 3: rating the quality of evidence. *J Clin Epidemiol*. 2011;64(4):401-406.

6. *Cochrane Handbook for Systematic Reviews of Interventions*. Version 5.1.0 ed. Chichester, United Kingdom: John Wiley & Sons Ltd; 2011.

7. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol*. 2009;62(10):1006-1012.

8. Murad MH, Montori VM. Synthesizing evidence: shifting the focus from individual studies to the body of evidence. *JAMA*. 2013;309(21):2217-2218.

9. Hopewell S, Clarke M, Lefebvre C, Scherer R. Handsearching versus electronic searching to identify reports of randomized trials. *Cochrane Database Syst Rev*. 2007;(2):MR000001.

10. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? empirical study. *Health Technol Assess*. 2003;7(1):1-76.

11. Livingston EH, Elliot A, Hynan L, Cao J. Effect size estimation: a necessary component of statistical analysis. *Arch Surg*. 2009;144(8):706-712.

12. Devereaux PJ, Yang H, Yusuf S, et al; POISE Study Group. Effects of extended-release metoprolol succinate in patients undergoing non-cardiac surgery (POISE trial): a randomised controlled trial. *Lancet*. 2008;371(9627):1839-1847.

13. Murad MH, Montori VM, Walter SD, Guyatt GH. Estimating risk difference from relative association measures in meta-analysis can infrequently pose interpretational challenges. *J Clin Epidemiol*. 2009; 62(8):865-867.

14. Guyatt GH, Eikelboom JW, Gould MK, et al; American College of Chest Physicians. Approach to outcome measurement in the prevention of thrombosis in surgical and medical patients: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2012;141(2)(suppl):e185S-e194S.

15. Taylor F, Huffman MD, Macedo AF, et al. Statins for the primary prevention of cardiovascular disease. *Cochrane Database Syst Rev*. 2013;1: CD004816.

16. Lacasse Y, Martin S, Lasserson TJ, Goldstein RS. Meta-analysis of respiratory rehabilitation in chronic obstructive pulmonary disease: a Cochrane systematic review. *Eura Medicophys*. 2007;43(4): 475-485.

17. Johnston BC, Patrick DL, Thorlund K, et al. Patient-reported outcomes in meta-analyses-part 2: methods for improving interpretability for decision-makers. *Health Qual Life Outcomes*. 2013; 11:211.

18. US Preventive Services Task Force (USPSTF). Grade Definitions. USPSTF website. http://www.uspreventiveservicestaskforce.org/uspstf/grades.htm. Accessed May 23, 2014.

19. Jacobs AK, Kushner FG, Ettinger SM, et al. ACCF/AHA clinical practice guideline methodology summit report: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. *Circulation*. 2013; 127(2):268-310.

20. Oxford Centre for Evidence-based Medicine (OECBM) Levels of Evidence Working Group. OECBM 2011 Levels of Evidence. OECBM website. http://www.cebm.net/mod_product/design/files/CEBM-Levels-of-Evidence-2.1.pdf. 2011. Accessed May 25, 2014.

21. Guyatt GH, Oxman AD, Vist GE, et al; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926.

22. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? *Lancet*. 1998;352(9128):609-613.

23. Odgaard-Jensen J, Vist GE, Timmer A, et al. Randomisation to protect against selection bias in healthcare trials. *Cochrane Database Syst Rev*. 2011; (4):MR000012.

24. Klein EA, Thompson IM Jr, Tangen CM, et al. Vitamin E and the risk of prostate cancer: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). *JAMA*. 2011;306(14):1549-1556.

25. Sesso HD, Buring JE, Christen WG, et al. Vitamins E and C in the prevention of cardiovascular disease in men: the Physicians' Health Study II randomized controlled trial. *JAMA*. 2008;300(18):2123-2133.

26. Jüni P, Witschi A, Bloch R, Egger M. The hazards of scoring the quality of clinical trials for meta-analysis. *JAMA*. 1999;282(11):1054-1060.

27. Higgins JP, 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.

28. Hatala R, Keitz S, Wyer P, Guyatt G; Evidence-Based Medicine Teaching Tips Working

Group. Tips for learners of evidence-based medicine, 4: assessing heterogeneity of primary studies in systematic reviews and whether to combine their results. *CMAJ*. 2005;172(5):661-665.

**29**. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127(9):820-826.

**30**. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-560.

**31**. Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ*. 2003; 326(7382):219.

**32**. Navarese EP, Kowalewski M, Andreotti F, et al. Meta-analysis of time-related benefits of statin therapy in patients with acute coronary syndrome undergoing percutaneous coronary intervention. *Am J Cardiol*. 2014;113(10):1753-1764.

**33**. Guyatt GH, Oxman AD, Kunz R, et al; GRADE Working Group. GRADE guidelines, 7: rating the quality of evidence—inconsistency. *J Clin Epidemiol*. 2011;64(12):1294-1302.

**34**. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines, 6: rating the quality of evidence—imprecision. *J Clin Epidemiol*. 2011;64 (12):1283-1293.

**35**. Dunkelgrun M, Boersma E, Schouten O, et al; Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. Bisoprolol and fluvastatin for the reduction of perioperative cardiac mortality and myocardial infarction in intermediate-risk patients undergoing noncardiovascular surgery: a randomized controlled trial (DECREASE-IV). *Ann Surg*. 2009; 249(6):921-926.

**36**. Poldermans D, Boersma E, Bax JJ, et al; Dutch Echocardiographic Cardiac Risk Evaluation Applying Stress Echocardiography Study Group. The effect of bisoprolol on perioperative mortality and myocardial infarction in high-risk patients undergoing vascular surgery. *N Engl J Med*. 1999; 341(24):1789-1794.

**37**. Murad MH, Shah ND, Van Houten HK, et al. Individuals with diabetes preferred that future trials use patient-important outcomes and provide pragmatic inferences. *J Clin Epidemiol*. 2011;64(7): 743-748.

**38**. Guyatt GH, Oxman AD, Kunz R, et al; GRADE Working Group. GRADE guidelines, 8: rating the quality of evidence—indirectness. *J Clin Epidemiol*. 2011;64(12):1303-1310.

**39**. Murad MH, Drake MT, Mullan RJ, et al. Comparative effectiveness of drug treatments to prevent fragility fractures: a systematic review and network meta-analysis. *J Clin Endocrinol Metab*. 2012;97(6):1871-1880.

**40**. Mills EJ, Ioannidis JP, Thorlund K, Schünemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. *JAMA*. 2012;308(12):1246-1253.

**41**. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337(8746):867-872.

**42**. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315(7109):640-645.

**43**. Chan AW, Hróbjartsson A, Haahr MT, Gøtzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*. 2004;291(20):2457-2465.

**44**. Saquib N, Saquib J, Ioannidis JP. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ*. 2013;347:f4313.

**45**. McDonagh MS, Peterson K, Balshem H, Helfand M. US Food and Drug Administration documents can provide unpublished evidence relevant to systematic reviews. *J Clin Epidemiol*. 2013;66(10):1071-1081.

**46**. Eyding D, Lelgemann M, Grouven U, et al. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ*. 2010;341:c4737.

**47**. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ*. 2006; 333(7568):597-600.

**48**. Boissel JP, Haugh MC. Clinical trial registries and ethics review boards: the results of a survey by the FICHTRE project. *Fundam Clin Pharmacol*. 1997; 11(3):281-284.

**49**. Horton R, Smith R. Time to register randomised trials: the case is now unanswerable. *BMJ*. 1999;319(7214):865-866.

**50**. Dickersin K, Rennie D. The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA*. 2012;307(17):1861-1864.

**51**. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines, 5: rating the quality of evidence—publication bias. *J Clin Epidemiol*. 2011; 64(12):1277-1282.

**52**. Guyatt GH, Oxman AD, Sultan S, et al; GRADE Working Group. GRADE guidelines, 9: rating up the quality of evidence. *J Clin Epidemiol*. 2011;64(12): 1311-1316.