Chapter 23: Understanding and Applying the Results of a Systematic Review and Meta-analysis

M. Hassan Murad; Victor M. Montori; John P. A. Ioannidis; Ignacio Neumann; Rose Hatala; Maureen O. Meade; PJ Devereaux; Peter Wyer; Gordon Guyatt

Introduction

In the previous chapter (Chapter 22, The Process of a Systematic Review and Meta-analysis), we provided guidance on how to evaluate the *credibility* of the process of a *systematic review* with or without a *meta-analysis*. In this chapter, we address how—if the systematic review is sufficiently credible—to decide on the degree of confidence in the estimates that the *evidence* warrants. As you will see, systematic review authors may have conducted a credible review and analysis and one may still have little confidence in the estimates of effect. We will return to the clinical scenario discussed in the previous chapter and obtain the relative and absolute effects of the intervention from a credible systematic review and meta-analysis[1] and determine the confidence in these estimates (quality of evidence). The general framework for judging confidence in estimates is based on the approach offered by the *GRADE* (*Grading of Recommendations Assessment, Development and Evaluation*) Working Group.[2] This chapter focuses on questions of therapy or harm. This framework can, however, be adapted for other types of questions, such as issues of prognosis[3] or diagnosis.[4]

CLINICAL SCENARIO

We continue with the scenario of a 66-year-old male smoker with type 2 diabetes and hypertension undergoing noncardiac surgery for whom we are considering prescribing perioperative β-blockers to prevent the cardiovascular complications of nonfatal infarction, death, and nonfatal stroke.

Understanding the Summary Estimate of a Meta-Analysis

If the systematic review authors decide that combining results to generate a single estimate of effect is inappropriate, a systematic review will likely end with a table or tables describing results of individual *primary studies*. Often, however, systematic reviews include a meta-analysis with a best estimate of effect (often called a summary or *pooled estimate*) from the weighted averages of the results of the individual studies. The weighting process depends on sample size or number of events (see Chapter 12.3, What Determines the Width of the Confidence Interval?) or, more specifically, study precision. Studies that are more precise have narrower *confidence intervals* (CIs) and larger weight in meta-analysis.

In a meta-analysis of a therapeutic question looking at *dichotomous outcomes* (yes/no) for estimates of the magnitude of the benefits or *risks*, you should look for the *relative risk* (RR) and *relative risk reduction* (RRR) or the *odds ratio* (OR) and *relative odds* reduction (see Chapter 9, Does Treatment Lower Risk? Understanding the Results). When the outcome is analyzed using time-to-event methods (eg, *survival analysis*), the results could be presented as a *hazard ratio*. In a meta-analysis addressing diagnosis, you should look for summary estimates of *likelihood ratios* or diagnostic ORs (see Chapter 18, Diagnostic Tests).

In the setting of *continuous variables* rather than dichotomous outcomes, meta-analysts typically use 1 of 2 options to aggregate data across studies. If the outcome is measured the same way in each study (eg, duration of hospitalization), the results from each study are combined, taking into account each study's precision to calculate what is called a *weighted mean difference*. This measure has the same units as the outcomes reported in the individual studies (eg, pooled estimate of reduction in hospital stay with treatment, 1.1 days).

Sometimes the outcome measures used in the primary studies are similar but not identical. For example, one trial might measure *health-related quality of life* using a validated questionnaire (the Chronic Respiratory Questionnaire), and another trial might use a different validated questionnaire (the St. George's Questionnaire). Another example of this situation is a meta-analysis of studies using different measures of severity of depression.

If the patients and the interventions are similar, generating a pooled estimate of the effect of the intervention on quality of life or depression, even when investigators have used different measurement instruments, is likely to be worthwhile. One way of generating the pooled estimate in this instance is to standardize the measures by looking at the mean difference between treatment and control and dividing this by the SD.[5] The *effect size* that results from this calculation provides a summary estimate of the treatment effect expressed in SD units (eg, an effect size of 0.5 means that the mean effect of treatment across studies is half of an SD unit). A rule of thumb for understanding effect sizes suggests that 0.2 SD represents small effects; 0.5 SD, moderate effects; and 0.8 SD, large effects.[6]

Clinicians may be unfamiliar with how to interpret effect size, and systematic review authors may help you interpret the results by using one of a number of alternative presentations. One is to translate the summary effect size back into natural units.[7] For instance, clinicians may have become familiar with the significance of differences in walk test scores in patients with chronic lung disease. Investigators can then convert the effect size of a treatment on a number of measures of functional status (eg, the walk test and stair climbing) back into differences in walk test scores.[8]

Even better may be the translation of continuous outcomes into dichotomies: the proportion of patients who, for instance, have experienced an important reduction in pain, fatigue, or dyspnea. Methods of making such translations are increasingly well developed.[9,10] For examples on how systematic review authors can present results that are ready for clinical applications, see Chapter 22, The Process of a Systematic Review and Meta-analysis.

The results of a traditional meta-analysis are usually depicted in what is called a *forest plot* (Figures 23-1, 23-2, and 23-3). This forest plot shows the effect (ie, result) from every study; the *point estimate* is presented as a square with a size that is proportional to the weight of the study, and the CI is presented as a horizontal line. The solid line at 1.0 indicates no effect, and the dashed line is centered on the meta-analysis combined summary effect. The combined summary effect is usually presented as a diamond, with its width representing the CI for the combined effect. As the CI widens, uncertainty about the magnitude of effect increases; when the CI crosses no effect (RR or OR of 1.0), there is uncertainty about whether the intervention has any effect at all (see Chapter 10, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?).

FIGURE 23-1

Results of a Meta-analysis of the Outcomes of Nonfatal Infarction in Patients Receiving Perioperative β-Blockers

Abbreviations: BBSA, Beta Blocker in Spinal Anesthesia study; CI, confidence interval; DIPOM, Diabetic Postoperative Mortality and Morbidity trial; MaVS, Metoprolol after Vascular Surgery study; POBBLE, Perioperative β-blockade trial; POISE, PeriOperative ISchemic Evaluation trial.

Solid line indicates no effect. Dashed line is centered on meta-analysis pooled estimate.

Data are from Bouri et al.[1]

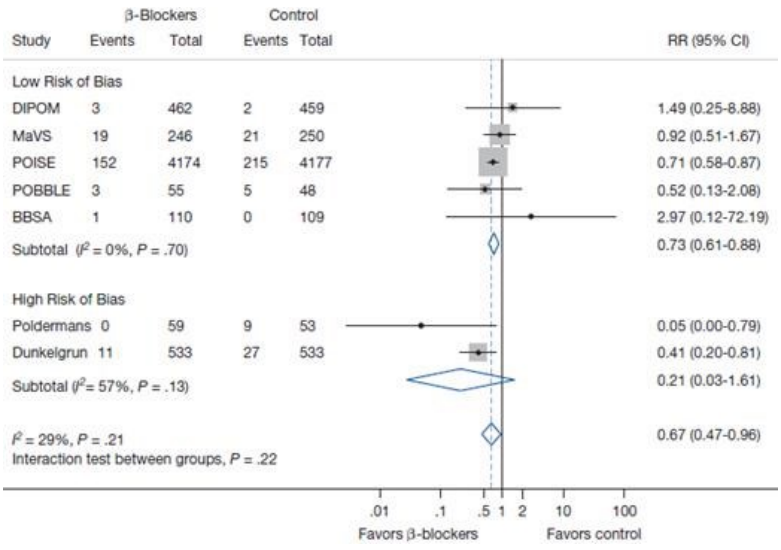


FIGURE 23-2

Results of a Meta-analysis of the Outcomes of Death in Patients Receiving Perioperative β-Blockers

Abbreviations: BBSA, Beta Blocker in Spinal Anesthesia study; CI, confidence interval; DIPOM, Diabetic Postoperative Mortality and Morbidity trial; MaVS, Metoprolol after Vascular Surgery study; POBBLE, Perioperative β-blockade trial; POISE, PeriOperative ISchemic Evaluation trial.

Solid line indicates no effect. Dashed line is centered on meta-analysis pooled estimate.

Data from Bouri et al.1

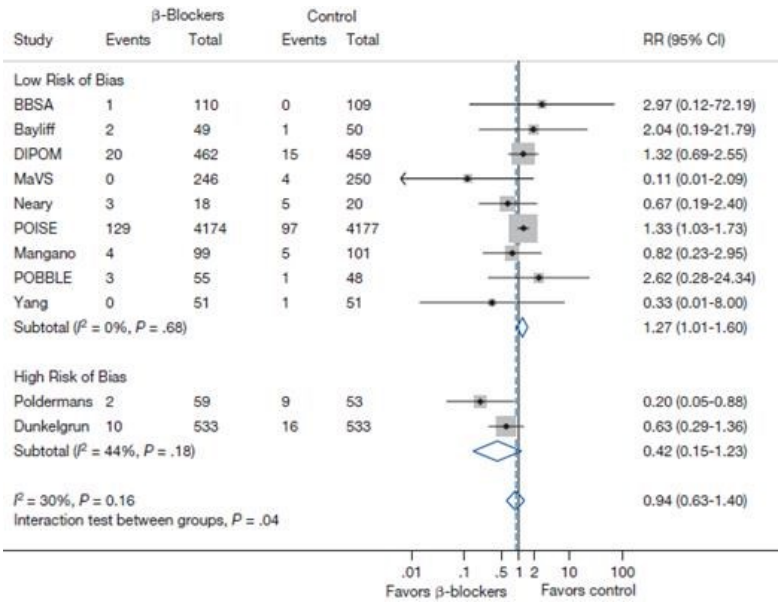


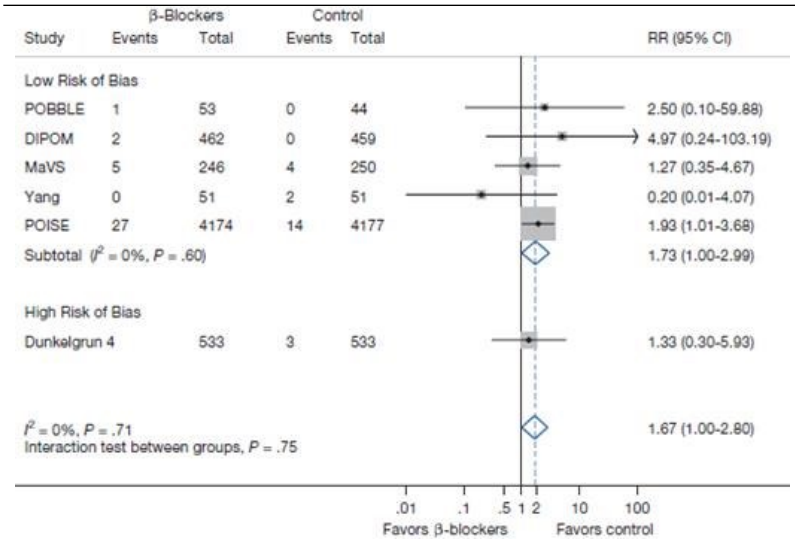
FIGURE 23-3

Results of a Meta-analysis of the Outcomes of Nonfatal Stroke in Patients Receiving Perioperative β-Blockers

Abbreviations: CI, confidence interval; DIPOM, Diabetic Postoperative Mortality and Morbidity trial; MaVS, Metoprolol after Vascular Surgery study; POBBLE, Perioperative β-blockade trial; POISE, PeriOperative ISchemic Evaluation trial.

Solid line indicates no effect. Dashed line is centered on meta-analysis pooled estimate.

Data from Bouri et al.1



USING THE GUIDE

Returning to the perioperative β-blockers scenario, you found a systematic review that you considered as having a credible process that included a meta-analysis for the outcomes of nonfatal infarction, mortality, and nonfatal stroke.[1] The forest plots reveal the estimates of effect for these outcomes from the relevant *randomized trials* (Figures 23-1, 23-2, and 23-3).

Perioperative administration of β-blockers decreases the risk of 1 adverse outcome—nonfatal myocardial infarction (RR, 0.67; 95% CI, 0.47-0.96). The summary effect reached the threshold for *statistical significance* because the CI does not cross 1.0 (no effect) (Figure 23-1). However, β-blockers likely increased the risk of nonfatal stroke, the lower boundary of the CI just touching no effect (RR, 1.67; 95% CI, 1.00-2.80) (Figure 23-3). You are not sure about the effect of β-blockers on the outcome of death because the CI crosses 1.0 and is wide, including a large reduction (37%) and a large increase (40%) in death (RR, 0.94; 95% CI, 0.63-1.40) (Figure 23-2).

You note, however, that there is appreciable *inconsistency* in the results for the *end points* of death and myocardial infarction and that, in particular, the studies with low or high *risk of bias* studies yield different results. This raises the question of which results are more credible, an issue to which we return later in this chapter.

Understanding the Estimate of Absolute Effect

The goal of a systematic review and meta-analysis is often to present evidence users (clinicians, patients, and policymakers) with best estimates of the effect of an intervention on each *patient-important outcome*. When interpreting and applying the results, you and your patients must balance the desirable and undesirable consequences to decide on the best course of action.

As we pointed out in the previous chapter (Chapter 22, The Process of a Systematic Review and Meta-analysis), knowledge of the RRs associated with the intervention is insufficient for making a decision about the trade-off between desirable and undesirable consequences; rather, it requires knowledge of the *absolute risk* associated with the intervention. For instance, the relative estimates we have presented so far suggest an RRR of myocardial infarction of 33% with use of β-blockers in noncardiac surgery but an increase in nonfatal strokes of 67%. The decision about whether to use β-blockers will be different, depending on whether the reduction in myocardial infarction is from 10% to 7% or from 1% to 0.7% and whether the increase in nonfatal strokes is from 0.5% to 0.8% or from 5% to 8%.

However, before we arrive at the best estimates of absolute effect we need to resolve a pending question: does the most trustworthy estimate of relative effect come from all of the studies or does it come from the studies with low risk of bias? We resolve this issue and present the best estimates of absolute effect later in this chapter.

Rating Confidence in the Estimates (The Quality of Evidence)

Consistent with the second principle of *evidence-based practice*—some evidence is more trustworthy and some less so—application of evidence requires a rating of how confident we are in our estimates of the magnitude of intervention effects on the outcomes of interest. This confidence rating is important for *clinical practice guideline* developers when they make their recommendations and for clinicians and patients when they decide on their course of action (see Chapter 26, How to Use a Patient Management Recommendation: Clinical Practice Guidelines and Decision Analyses, and Chapter 28.1, Assessing the Strength of Recommendations: The GRADE Approach).

The judgment about our confidence in the effect estimates applies not to a single study but rather a body of evidence. For any management decision, confidence in estimates can differ across outcomes. Historically, the word "quality" has been used synonymously with both risk of bias and confidence in estimates. Because of the ambiguity, we avoid the use of the word "quality" (although when we do use it, it is synonymous with confidence). Instead, we use the other 2 terms (risk of bias and confidence in estimates). In this chapter, the focus is on confidence in effect estimates.

The Grade Approach

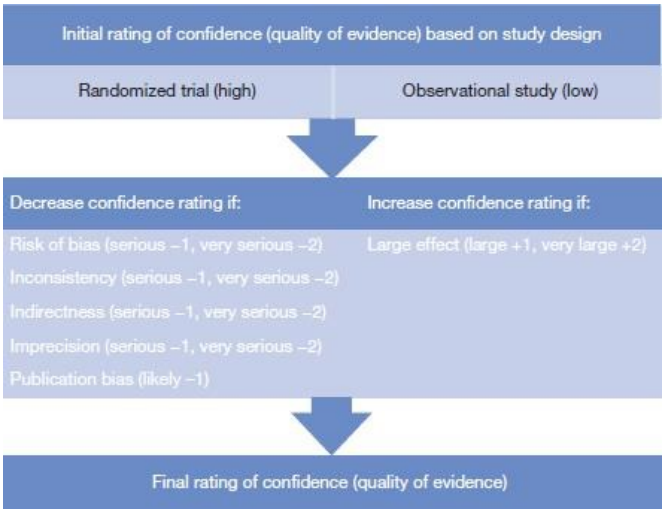
The GRADE approach is one of several systems to rate the quality of evidence. The GRADE Working Group is a group of health care professionals, researchers, and guideline developers who, in 2000, began to work together to develop an optimal system of rating confidence in estimates for systematic reviews and health technology assessments of questions of the impact of interventions and to determine the strength of recommendations for clinical practice guidelines.[2] The GRADE approach has been disseminated widely and endorsed by more than 70 organizations worldwide,[11,12] including the Cochrane Collaboration, the UK National Institutes of Clinical Excellence, the World Health Organization, and the American College of Physicians. Several hundred publications have since described, demonstrated the feasibility and usefulness, evaluated the use of, and provided guidance on the GRADE approach.

GRADE suggests rating confidence in estimates of effect in 4 categories: high, moderate, low, or very low. Some organizations, including UpToDate, combine the low and very low. The lower the confidence, the more likely the underlying true effect is substantially different from the observed estimate of effect, and thus, it is more likely that further research would reveal a change in the estimates.[13]

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 (Figure 23-4). Confidence ratings may decrease when there is increased risk of bias, inconsistency, *imprecision*, *indirectness*, or concern about *publication bias.* An increase in confidence rating is uncommon and mainly occurs when the effect size is large (Figure 23-4).

FIGURE 23-4

Rating the Quality of Evidence (the Confidence in the Estimates)



These factors defined by GRADE should affect our confidence in estimates whether or not systematic review authors formally use GRADE. In one way or

another, therefore, your consideration of evidence from a systematic review of alternative management strategies must include consideration of these issues. We now provide a description of how authors of systematic reviews and meta-analyses apply these criteria.

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

Authors of systematic reviews evaluate the risk of bias for each of the outcomes measured in each individual study. *Bias* represents systematic rather than random error (see Chapter 6, Why Study Results Mislead: Bias and Random Error).

For randomized trials, risk of bias increases if there are problems with the *randomization* (defects in generation of the randomization sequence or lack of appropriate *allocation concealment*); if patients, caregivers, and study personnel are not *blinded;* or if a large number of patients are *lost to follow-up* (see Chapter 7, Therapy [Randomized Trials]). The effect of these problems can differ across outcomes. For example, lack of blinding and inadequate allocation concealment lead to greater bias for subjective outcomes than for objective hard clinical outcomes, such as death.[14] Stopping trials early because of a large apparent effect also can exaggerate the treatment effects (see Chapter 11.3, Randomized Trials Stopped Early for Benefit).[15] In observational studies, the main concerns associated with increased risk of bias include inappropriate measurement of *exposure* and outcome, inadequate statistical adjustment for prognostic imbalance, and loss to follow-up (see Chapter 14, Harm [Observational Studies]).

Ideally, the authors of systematic reviews will present a risk of bias evaluation for every individual study and provide a statement about the overall risk of bias for all of the included studies. The reproducibility of this judgment affects the credibility of the process of the systematic review (see Chapter 22, The Process of a Systematic Review and Meta-analysis). Following the GRADE approach, the risk of bias can be expressed as "not serious," "serious," or "very serious." The assessment of the level of risk of bias can then result in no decrease in the confidence rating in estimates of effect or a decrease by 1 or 2 levels (eg, from high to moderate or low confidence) (Figure 23-4).[13]

USING THE GUIDE

The authors of the systematic review and meta-analysis addressing perioperative β-blockers[1] used the Cochrane Collaboration risk of bias assessment methods (see Chapter 22, The Process of a Systematic Review and Meta-analysis). They explicitly described the risk of bias of each trial and reported on the adequacy of generation of the allocation sequence; allocation concealment; blinding of participants, personnel, and outcome assessors; the extent of loss to follow-up; and the use of the *intention-to-treat principle*.

Of the 11 trials included in the analysis, 2 were considered to have high risk of bias[16,17]; limitations included lack of blinding and, in one trial, *stopping early* because of large apparent benefit.[17] The results of these 2 trials became even more questionable when, subsequently, concerns were raised about the integrity of the data.[1] The remaining 9 trials were deemed by the systematic review authors to have adequate bias protection measures and represented a body of evidence that was overall at low risk of bias for the 3 key outcomes—nonfatal myocardial infarction, death, and nonfatal stroke.

Are the Results Consistent Across Studies?

The starting assumption of a meta-analysis that provides a summary estimate of treatment effect is that across the range of study patients, interventions, and outcomes included in the analysis, the effect of interest is more or less the same (see Chapter 22, The Process of a Systematic Review and Meta-analysis). On the one hand, a meta-analysis question framed to include a broad range of patients, interventions, and ways of measuring outcome helps avoid spurious effects from *subgroup analyses* (see Chapter 25.2, How to Use a Subgroup Analysis), leads to narrower CIs, and increases applicability across a broad range of patients. On the other hand, combining the results of diverse studies may violate the starting assumption of the analysis and lead to spurious conclusions (for instance, that the same estimate of effect applies to different patient groups or different ways of administering an intervention, when it in fact does not).

The solution to this dilemma is to evaluate the extent to which results differ from study to study, that is, the variability or *heterogeneity* of study results. Box 23-1 summarizes 4 approaches to evaluating variability in study results, and the subsequent discussion expands on these principles.[18]

BOX 23-1

Evaluating Variability in Study Results

- Visual evaluation of variability
  - How similar are the point estimates?
  - To what extent do the confidence intervals overlap?
- Statistical tests evaluating variability
  - Yes-or-no tests for heterogeneity that generate a *P* value
  - I^2 test that quantifies the variability explained by between-study differences in results

Visual Assessment of Variability

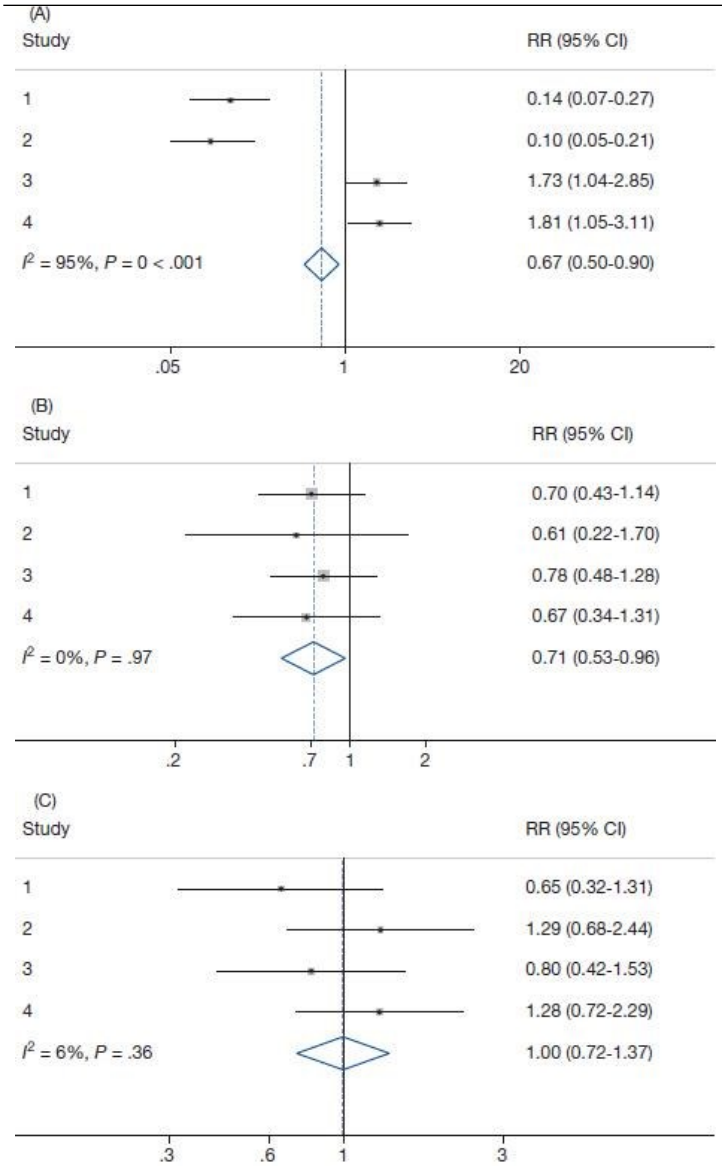
Studies combined in a meta-analysis and depicted in a forest plot will inevitably have some inconsistency (heterogeneity) of their point estimates. The question is whether that heterogeneity is sufficiently great to make us uncomfortable with combining results from a group of related studies to generate a single summary effect.[19]

Consider the results of the 2 meta-analyses shown in Figures 23-5A and B (meta-analysis A and meta-analysis B, respectively). When reviewing the results of these studies, would clinicians be comfortable with a single summary result in either or both meta-analyses? Although the results of meta-analysis A seem extremely unlikely to meet the assumption of a single underlying treatment effect across studies, the results of meta-analysis B are completely consistent with the assumption. Therefore, we would be uncomfortable applying the pooled estimate to all studies in A but comfortable doing so in B.

FIGURE 23-5

Results of Hypothetical Meta-analyses

Abbreviations: CI, confidence interval; RR, relative risk.



Constructing a rule to capture these inferences, one might suggest that "we are comfortable with a single summary effect when all studies suggest benefit or all studies suggest harm" (the case for B but not A). Figure 23-5C, however, highlights the limitation of such a rule: this hypothetical meta-analysis C also shows point estimates on both sides of the line of no effect, but here we would be comfortable combining the results.

A better approach to assessing heterogeneity focuses on the magnitude of the differences in the point estimates of the studies. Large differences in point estimates make clinicians less confident in the pooled estimate (as in meta-analysis A). Small differences in the magnitude of point estimates (as in meta-analyses B and C) support the underlying assumption that, across the range of study patients, interventions, and outcomes included in the meta-analysis, the effect of interest is more or less the same.

There is a second, equally important criterion that clinicians should apply when judging whether combining the studies is appropriate. If CIs overlap widely (as in meta-analyses B and C), *random error*, or chance, remains a plausible explanation for the differences in the point estimates. When CIs do not overlap (as in meta-analysis A), random error becomes an unlikely explanation for differences in apparent treatment effect across studies. Visual assessment of heterogeneity is useful; formal statistical testing can provide complementary information.

Yes-or-No Statistical Tests of Heterogeneity

The *null hypothesis* (see Chapter 12.1, Hypothesis Testing) of the *test for heterogeneity* is that the underlying effect is the same in each of the studies[20]

(eg, the RR derived from study 1 is the same as that from studies 2, 3, and 4). Therefore, the null hypothesis assumes that all of the apparent variability among individual study results is due to chance. *Cochran Q*, the most commonly used test for heterogeneity, generates a *probability* based on a χ^2 distribution that between-study differences in results equal to or greater than those observed are likely to occur simply by chance.

Meta-analysts may consider different thresholds for the significance of the test of heterogeneity (eg, a conventional threshold of $P < .05$ or a more conservative threshold of $P < .10$). As a general principle, however, a low P value of the test for heterogeneity means that random error is an unlikely explanation for the differences in results from study to study. Thus, a low P value decreases confidence in a single summary estimate that represents the treatment effect for all patients and all variations in the administration of a treatment. A high P value of the test of heterogeneity, on the other hand, increases our confidence that the assumption underlying combining studies holds true.

In Figure 23-5A, the P value associated with the test for heterogeneity is small ($P < .001$), indicating that it is unlikely that we would observe results this disparate if all studies had the same underlying effect. On the other hand, the corresponding P values in Figure 23-3B and C are fairly large (.97 and .36, respectively). Therefore, in these 2 meta-analyses, chance is a likely explanation for the observed differences in effect.

When a meta-analysis includes studies with small sample sizes and a correspondingly small number of events, the test of heterogeneity may not have sufficient power to detect existing heterogeneity. Conversely, in a meta-analysis that includes studies with large sample sizes and a large number of events, the test for heterogeneity may provide potentially misleading results that reveal statistically significant but unimportant differences in point estimates. This is another reason why clinicians need to use their own visual assessments of heterogeneity (similarity of point estimates, overlap of CIs) and consider the results of formal statistical tests in that context.

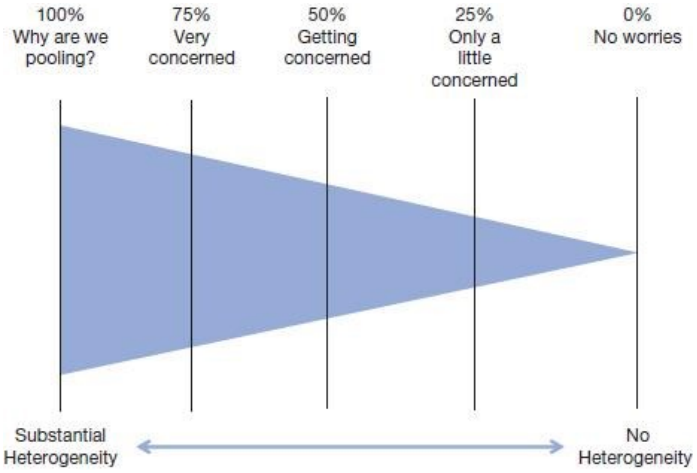
Magnitude of Heterogeneity Statistical Tests

The I^2 *statistic* is a preferred alternative approach for evaluating heterogeneity that focuses on the magnitude of variability rather than the statistical significance of variability.[21]

When the I^2 is 0%, chance provides a satisfactory explanation for the variability in the individual study point estimates, and clinicians can be comfortable with a single summary estimate of treatment effect. As the I^2 increases, we become progressively less comfortable with a single summary estimate, and the need to look for explanations of variability other than chance becomes more compelling. Figure 23-6 provides a guide for interpreting the I^2.

FIGURE 23-6

Interpretation of the I^2 Statistic



If provided by the meta-analysis authors, a 95% CI associated with the I^2 can provide further insight regarding assessment of inconsistency. In most meta-analyses with a limited number of relatively small studies, this CI is quite large, suggesting the need for caution in making strong inferences regarding inconsistency.[22]

The results in Figure 23-5A generate an I^2 of more than 75% (suggesting high heterogeneity), whereas the results in Figure 23-5B and C yield low I^2 percentages of 0% and 6%, respectively (suggesting low heterogeneity).

What to Do When Between-Study Variability in Results Is Large?

One of the credibility criteria introduced in Chapter 22 is whether the authors have addressed possible explanations of heterogeneity. When between-study variability is large, such an exploration becomes crucial.

Differences between study results can arise from differences in the population enrolled (eg, large effects in the more ill, smaller in the less ill), differences in the interventions (eg, if large doses are more effective than small doses), differences in the comparators (eg, smaller effects when standard care is optimal than when it is not), and study methods (eg, larger effect in studies with high risk of bias vs those with low risk of bias). Meta-analysis authors should conduct a test of interaction to determine whether the difference in effect estimates among subgroups is attributable to chance. Apparent subgroup effects are more likely to be true when they are based on within-trial rather than between-trial comparisons, are very unlikely to be due to chance, and are based on a small number of hypotheses specified a priori, including a specified direction. If these criteria are not met, any subgroup hypothesis warrants a high level of skepticism (see Chapter 25.2, How to Use a Subgroup Analysis). For a discussion of additional issues in statistical analysis related to heterogeneity of study results, see Chapter 25.1, Fixed-Effects and Random-Effects Models.

What if, in the end, we are left with a large degree of unexplained between-study heterogeneity for which chance does not provide an adequate explanation? This is not an uncommon situation. Some argue that, in this situation, meta-analysis authors should not combine the results. Clinicians and patients, however, still need a best estimate of the treatment effect to inform their decisions. Pending further research that may explain the differences between results of different studies that address the same question, the summary estimate remains the best estimate of the treatment effect. Although clinicians and patients must use the best estimate to make their decisions, substantial unexplained inconsistency between studies appreciably reduces confidence in the summary estimate.[23]

USING THE GUIDE

In Figures 23-1 and 23-2, for both nonfatal myocardial infarction and death, we note substantial differences in point estimates across studies. In the case of death, there is minimal CI overlap. Although the heterogeneity *P* values of .21 and .16 are not statistically significant, the I^2 of 29% for nonfatal myocardial infarction and 30% for death suggest the presence of variability for which seeking a possible explanation is worthwhile.

Examining the data, we find that trials with a high risk of bias reveal a substantially larger reduction in the risk of nonfatal myocardial infarction. A test of interaction between the 2 groups of studies (those with high risk of bias and those with low risk of bias) yields a nonsignificant *P* value of .22, which indicates that the difference in the reduction in nonfatal myocardial infarction risk between these 2 subgroups of studies could be due to chance.

However, for the outcome of death, a test of interaction between the 2 groups of studies yields a significant *P* value of .04, which suggests that the risk of bias explains the observed heterogeneity (Figure 23-2). As we have mentioned previously, our inclination to use only the studies with low risk of bias is reinforced by our awareness of the doubts that have been raised regarding the integrity of the data from the 2 studies with high risk of bias. Results of the studies with low risk of bias are consistent (I^2 of 0% and *P* value for heterogeneity test of .68).

Meta-analysis of the outcome of nonfatal stroke reveals consistent results across trials with an I^2 value of 0% and *P* value for the heterogeneity test of .71 (Figure 23-3).

How Precise Are the Results?

Meta-analysis generates an estimate of the mean effect across studies and a CI around that estimate, that is, a range of values with a specified probability (typically 95%) of including the true effect (see Chapter 10, Confidence Intervals: Was the Single Study or Meta-analysis Large Enough?). When applying research evidence to a clinical question, one should determine whether clinical action would differ if the upper or the lower boundaries of the CI represented the truth. If the clinical decision is the same whether the upper or lower boundary of the CI represents the true effect, then the

evidence is sufficiently precise. If across the range of the CI values our decision making would change, then we should have less confidence in the evidence and lower the confidence rating (eg, from high to moderate confidence).[24]

USING THE GUIDE

To determine the precision of the estimate of the effect of perioperative β-blockers on the risk of nonfatal myocardial infarction, you need to calculate the absolute effect, which requires knowledge of the RR and the *control event rate* (ie, the event rate in patients who did not receive β-blockers). Having decided that the best estimate of RR comes from focusing on the trials with low risk of bias rather than all trials included in the meta-analysis, we note that the RR is 0.73 (95% CI, 0.61-0.88) (Figure 23-1). We obtain the control event rate from the trial that is by far the largest—and the one that likely enrolled the most representative population[25]—which was 215/4177 or approximately 52 per 1000. You can then calculate the decreased risk of nonfatal myocardial infarction in those using β-blockers as follows:

Risk with intervention = risk with control × relative risk = 52/1000 × 0.73 = approximately 38 per 1000

Risk difference = risk with control − risk with intervention = 52/1,000 − 38/1000 = −14 (approximately 14 fewer myocardial infarctions per 1000)

You can use the same process to calculate the CIs around the *risk difference*, substituting the boundaries of the CI (in this case, 0.61 and 0.88) for the point estimate (in this case, 0.73). For instance, for the upper boundary of the CI:

Risk with intervention = 52/1000 × 0.88 = approximately 46 per 1000

Risk with intervention − risk with control = 46 − 52 = −6 (approximately 6 fewer per 1000)

The estimate of absolute difference in nonfatal myocardial infarction when using β-blockers is therefore approximately 14 fewer per 1000, with a CI of approximately 6 to 20 fewer per 1000.

The corresponding absolute difference for nonfatal stroke is 2 more nonfatal strokes per 1000, with a CI of approximately 0 to 6 more per 1000; for death, the absolute difference is 6 deaths more per 1000 with a CI of approximately 0 to 13 more per 1000 (Table 23-1).

Lowering a confidence rating because of imprecision is always a judgment call. There seems to be no doubt about the need to lower confidence for nonfatal stroke (the effect ranges from no difference to an appreciable increase in nonfatal stroke) and likely for death (some may consider 1 additional death in 1000 acceptable given the reduction in myocardial infarction; most would not consider 6 in 1000 trivial). Regarding nonfatal myocardial infarction, our judgment was not to lower confidence for imprecision (Table 23-1).

TABLE 23-1

Evidence Profile: Explicit Presentation of the Best Estimates of the Effect of Perioperative β-Blockers and the Confidence in Estimates

Confidence Assessment			
Outcome	Myocardial infarction	Stroke	Death
No. of Participants (No. of Studies)	10 189 (5)	10 186 (5)	10 529 (9)
Risk of Bias	No serious limitations	No serious limitations	No serious limitations
Consistency	No serious limitations	No serious limitations	No serious limitations
Directness	No serious limitations	No serious limitations	No serious limitations
Precision	No serious limitations	Imprecise	Imprecise
Reporting Bias	Not detected	Not detected	Not detected
Summary of Findings			
Confidence	High	Moderate	Moderate
Relative Effect (95% CI)	0.73 (0.61-0.88)	1.73 (1.00-2.99)	1.27 (1.01-1.60)
Risk Difference per 1000 Patients	14 fewer (6 fewer to 20 fewer)	2 more (0 more to 6 more)	6 more (0 more to 13 more)

Abbreviation: CI, confidence interval.

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, and outcomes in studies differ from those of interest (ie, the patient before us), we lose confidence in estimates of effect. In GRADE, the term "indirectness" is used as a label for these issues.[26]

So, for instance, the patient at hand may be very elderly and the trials may have included few, if any, such patients. The dose of a drug tested in the trials may be greater than the dose your patient can tolerate.

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 (see Chapter 13.1, Applying Results to Individual Patients). Do elderly patients metabolize a drug differently from younger patients? Are there competing risks that will be responsible for the demise of elderly patients long before they experience the benefits of the intervention? Is there evidence that the tissue effect of a medication is highly dose dependent?

USING THE GUIDE

Assessing directness regarding the evidence bearing on the use of β-blockers in noncardiac surgery,[1] we note that the age of most patients enrolled across the trials ranged from 50 to 70 years, similar to your patient, who is 66 years old. Almost all of the trials enrolled patients undergoing surgical procedures classified as intermediate surgical risk, similar to the hip surgery of your patient. Most of the trials enrolled many patients who, like yours, had risk factors for heart disease. Although the drug used and the dose varied across trials, the consistent results suggest you can use a modest dose of the β-blocker with which you are most familiar. The outcomes of death, nonfatal stroke, and nonfatal infarction are the key outcomes of importance to your patients. Overall, the available evidence presented in the systematic review is direct and applicable to your patient and addresses the key outcomes (benefits and harms) needed for decision making.

Another issue of indirectness arises when outcomes assessed in the studies differ from those of interest to patients. Trials often measure laboratory or *surrogate outcomes* that are not themselves important but are measured in the presumption that changes in the surrogate reflect changes in an outcome important to patients (see Chapter 13.4, Surrogate Outcomes). For instance, we have excellent information about the effect of medications used in type 2 diabetes on hemoglobin A_{1C} but limited information on their effect on macrovascular and microvascular disease. In almost every instance, we should reduce our confidence in estimates of effect on patient-important outcomes when all we have available is the effect on surrogates.

Lastly, a different type of indirectness occurs when clinicians must choose among interventions that have not been tested in head-to-head comparisons. For instance, we may want to choose among alternative bisphosphonates for managing osteoporosis. We will find many trials that compare each agent to placebo, but few, if any, that have compared them directly against one another.[27] Making comparisons among treatments under these circumstances requires extrapolating results for the existing comparisons and requires multiple assumptions (see Chapter 24, Network Meta-analysis).[26]

Is There Concern About Reporting Bias?

The most difficult types of bias for systematic review authors to address stem from the inclination of authors of original studies to publish material, either entire studies or specific outcomes, based on the magnitude, direction, or statistical significance of the results. We call the systematic error in the body of evidence that results from this inclination *reporting bias*. When an entire study remains unreported, the standard term is publication bias. The reason for publication bias is that studies without statistically significant results (*negative studies*) are less likely to be published than studies that reveal apparent differences (*positive studies*). The magnitude and direction of a study's results may be more important determinants of publication than study design, relevance, or quality,[28] and positive studies may be as much as 3 times more likely to be published than negative studies.[29] When authors or study sponsors selectively manipulate and report specific outcomes and analyses, we use the term *selective outcome reporting bias*.[30] Selective reporting bias can be a serious problem. Empirical evidence suggests that half of the analysis plans of randomized trials are different in protocols than in published reports.[31] When the publication is delayed because of the lack of significance of results, authors have used the term *time lag bias*.[32]

Selective outcome reporting can also create misleading estimates of effect. A study of US Food and Drug Administration (FDA) reports found that they often included numerous unpublished studies and the findings of these studies can appreciably alter the estimates of effect.[33]

Reporting bias can intrude at virtually all stages of the planning, implementation, and dissemination of research. Even if studies with negative results succeed and get published, they may still suffer from *dissemination bias*: they may be published in less prominent journals, may not receive adequate attention from policymakers, may be omitted (whether identified or not) in narrative reviews, may be omitted (if unidentified) from systematic reviews, and may have minimal or no effect on formulation of policy guidelines. On the other hand, studies with positive results may receive disproportionate attention. For instance, they are more likely to appear in subsequent evidence summaries and in an evidence *synopsis*.[34]

The consequences of publication and reporting bias can corrupt the body of evidence, usually exaggerating estimates of magnitude of *treatment effect*. Systematic reviews that fail to identify and include unpublished studies face a risk of presenting overly sanguine estimates of treatment effectiveness.

The risk of publication bias is probably higher for systematic reviews and meta-analyses that are based on small studies. Small studies are more likely to produce nonsignificant results due to lack of statistical power and are easier to hide. Larger studies are not, however, immune. Sponsors and authors who are not pleased with the results of a study may delay publication or chose to publish their study in a journal with limited readership or a lower impact factor.[32]

An example of reporting bias is the Salmeterol Multicenter Asthma Research Trial, which was a randomized trial designed to examine the effect of salmeterol or placebo on a *composite end point* of respiratory-related deaths and life-threatening experiences. In September 2002, after a data safety and monitoring board review of 25 858 randomized patients that found a nearly significant increase in the primary outcome in salmeterol-treated patients, the sponsor terminated the study. In a significant deviation from the original protocol, the sponsor submitted to the FDA an analysis, including events in the 6 months after the termination of the trial, which produced an apparent diminution of the dangers associated with salmeterol. The FDA, through specific inquiry, eventually obtained the data and the results were finally published in January 2006, revealing the increased likelihood of respiratory-related deaths with salmeterol.[35,36]

Strategies to Address Reporting Bias

Several tests have been developed to detect publication bias (Box 23-2); unfortunately, all have serious limitations. The tests require a large number of studies (ideally 30 or more), although many meta-analysis authors use them in analyses including few studies. Moreover, none of these tests has been validated against a *criterion standard* (or *gold standard*) of real data in which we know whether publication bias or other biases existed or not.[37]

BOX 23-2

Four Strategies to Address Reporting Bias

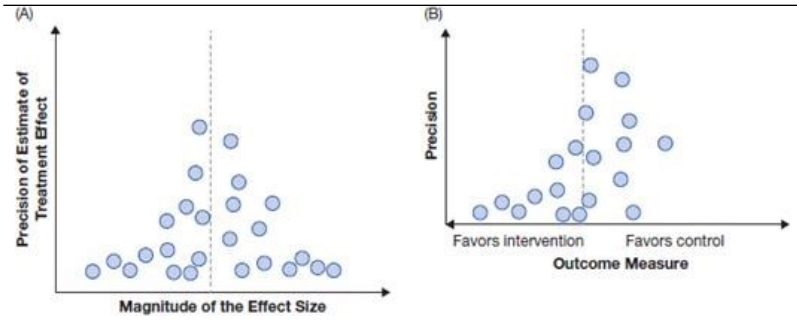
1. Examine whether the smaller studies show bigger effects
   a. Funnel plots, visually assessed
   b. Funnel plots, statistical analysis
2. Reconstruct evidence by restoring the picture after accounting for postulated publication bias
   a. Trim and fill
3. Estimate the chances of publication according to the statistical significance level
4. Examine the evolution of effect size over time as more data appear

The first category of tests examines whether small studies differ from larger ones in their results. In a figure that relates the precision (as measured by sample size, inverse of *standard error* or *variance*) of studies included in a meta-analysis to the magnitude of treatment effect, the resulting display should resemble an inverted funnel (Figure 23-7A). Such *funnel plots* should be symmetric, around the point estimate (dominated by the largest trials) or the results of the largest trials themselves. A gap or empty area in the funnel suggests that studies have been conducted and not published (Figure 23-7B). Because visual determination of symmetry can be subjective, meta-analysts sometimes apply statistical tests for the symmetry of the funnel.[37]

FIGURE 23-7

Funnel Plot Showing No Publication Bias (A) and Showing Possible Publication Bias (B)

A, 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 summary estimate (the dashed line). In this case, the effect sizes of the smaller studies are more or less symmetrically distributed around the summary estimate. B, 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. This suggests publication bias and an overestimate of the treatment effect relative to the underlying truth.



Even when the funnel shape or the tests suggest publication bias, other explanations for asymmetry are possible. The small studies may have a higher risk of bias, which may explain their larger effects. On the other hand, the small studies may have chosen a more responsive patient group or administered the intervention more meticulously. Finally, there is always the possibility of a chance finding.

A second set of tests imputes and corrects for missing information and address its effect (*trim-and-fill method*). Again, the availability of few studies and the presence of heterogeneity make this second strategy inappropriate for most meta-analyses.

A third set of tests estimates whether there are differential chances of publication according to the level of statistical significance.[38-40] The excess significance test can be used in single meta-analyses and collections of multiple meta-analyses in the same field where similar biases may be operating.

Finally, a set of tests aims to examine whether evidence changes over time as more data accumulate. Continuously diminishing effects are characteristic of time lag bias.[41]

More compelling than any of these theoretical exercises is the success of systematic review authors in obtaining the results of unpublished studies that appear to be a complete collection of all of the studies that have been undertaken.

Prospective study registration with accessible results represents the best solution to reporting bias.[42,43] Prospective registration makes publication bias potentially identifiable; however, more detailed information is necessary to identify potential selective outcome and analysis reporting bias. Until complete reporting becomes a reality,[44] clinicians using research reports to guide their practice must remain cognizant of the dangers of reporting biases.

USING THE GUIDE

The authors of the systematic review and meta-analysis[1] addressing perioperative β-blockers constructed funnel plots that appear to be symmetrical, and the statistical tests for the symmetry of the plot were nonsignificant. The total number of patients included (>10 000) further reduces concern about publication bias, leaving no reason for lowering our confidence rating due to publication or reporting bias.

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.[45] In each of these situations, we are confident of a substantial treatment effect despite the absence of randomized trials. Why is that? The reason is a very large treatment effect that was achieved during a short period among patients with a condition that would have inevitably worsened in the absence of an intervention.

The GRADE approach provides specific guidance regarding large effect sizes: consider increasing the confidence rating by 1 level when there is a 2-fold reduction or increase in risk and consider increasing the confidence rating by 2 levels in the presence of a 5-fold reduction or increase in risk. For example, a systematic review and meta-analysis of observational studies examining the relationship between infant sleeping position and sudden infant death syndrome (SIDS) found an OR of 4.9 (95% CI, 3.6-6.6) of SIDS occurring with front vs back sleeping positions.[46] The “back to sleep” campaigns that were started in the 1980s were associated with a relative decrease in the incidence of SIDS by 50% to 70% in numerous countries.[46]

This large effect increases our confidence in a true association.[45]

An Evidence-Based Summary of the Findings: The Evidence Profile

To optimally apply evidence summarized in a systematic review, practitioners need succinct, easily digestible presentations of confidence in effect estimates (quality of evidence) and magnitude of effects. They need this information to trade benefits and harms and communicate risks to their patients. They need to know the confidence we have in a body of evidence to convey the uncertainty to their patients.

Systematic reviews may provide this summary in different ways. The GRADE Working Group recommends what are called *evidence profiles* (or a shortened version called *summary of findings tables*). Such tables present the relative and absolute effects of an intervention on each of the critical outcomes most important to patients, including a confidence rating. If stratifying patients' baseline risk for the outcome is possible, the absolute effect is presented for each risk strata separately.

CLINICAL SCENARIO RESOLUTION

Table 23-1 presents the evidence profile summarizing the results of the systematic review addressing perioperative β-blockers. We see that evidence warranting high confidence suggests that individuals with underlying cardiovascular disease or risk factors for disease can expect a reduction in their 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 experiencing a nonfatal stroke. Because most people are highly averse to the disability associated with stroke and at least equally averse to death, it is likely that most patients faced with this evidence would decline β-blockers as part of their perioperative regimen. Indeed, that is what our 66-year-old man with diabetes decides when you discuss the evidence with him.

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. *Heart*. 2014:100(6):456-–464. [PubMed: 23904357]

2. 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. [PubMed: 18436948]

3. Spencer FA, Iorio A, You J, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. *BMJ*. 2012;345:e7401. [PubMed: 23152569]

4. Schünemann HJ, Oxman AD, Brozek J, et al; GRADE Working Group. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ*. 2008;336(7653):1106-–1110. [PubMed: 18483053]

5. Rosenthal R. *Meta-analytic Procedures for Social Research.* 2nd ed. Newbury Park, CA: Sage Publications; 1991.

6. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Earlbaum Associates; 1988.

7. Smith K, Cook D, Guyatt GH, Madhavan J, Oxman AD. Respiratory muscle training in chronic airflow limitation: a meta-analysis. *Am Rev Respir Dis*. 1992;145(3):533-–539. [PubMed: 1532118]

8. 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. [PubMed: 18084170]

9. Thorlund K, Walter S, Johnston B, Furukawa T, Guyatt G. Pooling health-related quality of life outcomes in meta-analysis—a tutorial and review of methods for enhancing interpretability. *Res Synth Methods*. 2011;2(3):188-–203. [PubMed: 26061786]

10. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol*. 2013;66(2):173-–183. [PubMed: 23116689]

11. Guyatt GH, Oxman AD, Schünemann HJ, Tugwell P, Knottnerus A. GRADE guidelines: a new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*. 2011;64(4):380-–382. [PubMed: 21185693]

12. Organizations. *The GRADE Working Group*. http://www.gradeworkinggroup.org/society/index.htm. Accessed April 9 , 2014.

13. Balshem H, Helfand M, Schünemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol*. 2011;64(4):401-–406. [PubMed: 21208779]

14. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*. 2008;336(7644):601-–605. [PubMed: 18316340]

15. Bassler D, Briel M, Montori VM, et al; STOPIT-2 Study Group. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*. 2010;303(12):1180-–1187. [PubMed: 20332404]

16. 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. [PubMed: 19474688]

17. 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. [PubMed: 10588963]

18. 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. [PubMed: 15738493]

19. Lau J, Ioannidis JP, Schmid CH. Summing up evidence: one answer is not always enough. *Lancet*. 1998;351(9096):123-–127. [PubMed: 9439507]

20. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. *Ann Intern Med*. 1997;127(9):820-–826. [PubMed: 9382404]

21. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003;327(7414):557-–560. [PubMed: 12958120]

22. Ioannidis JP, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007;335(7626):914-–916. [PubMed: 17974687]

23. 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. [PubMed: 21803546]

24. 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. [PubMed: 21839614]

25. 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. [PubMed: 18479744]

26. 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. [PubMed: 21802903]

27. Murad MH, Drake MT, Mullan RJ, et al. Clinical review. 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. [PubMed: 22466336]

28. Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. *Lancet*. 1991;337(8746):867-–872. [PubMed: 1672966]

29. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. 1997;315(7109):640-–645. [PubMed: 9310565]

30. Chan AW, Hróbjartsson A, Haahr MT, G⊘tzsche 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. [PubMed: 15161896]

31. Saquib N, Saquib J, Ioannidis JP. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ*. 2013;347:f4313. [PubMed: 23851720]

32. Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA*. 1998;279(4):281-–286. [PubMed: 9450711]

33. 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. [PubMed: 23856190]

34. Carter AO, Griffin GH, Carter TP. A survey identified publication bias in the secondary literature. *J Clin Epidemiol*. 2006;59(3):241-–245. [PubMed: 16488354]

35. Lurie P, Wolfe SM. Misleading data analyses in salmeterol (SMART) study. *Lancet*. 2005;366(9493):1261-–1262, discussion 1262. [PubMed: 16214589]

36. Nelson HS, Weiss ST, Bleecker ER, Yancey SW, Dorinsky PM; SMART Study Group. The Salmeterol Multicenter Asthma Research Trial: a comparison of usual pharmacotherapy for asthma or usual pharmacotherapy plus salmeterol. *Chest*. 2006;129(1):15-–26. [PubMed: 16424409]

37. Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. The case of the misleading funnel plot. *BMJ*. 2006;333(7568):597-–600. [PubMed: 16974018]

38. 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(4):299-–333.

39. Vevea J, Hedges L. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*. 1995;60(3):419-–435.

40. Ioannidis JP, Trikalinos TA. An exploratory test for an excess of significant findings. *Clin Trials*. 2007;4(3):245-–253. [PubMed: 17715249]

41. 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(4):281-–291. [PubMed: 10235168]

42. 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. [PubMed: 9243261]

43. Horton R, Smith R. Time to register randomised trials. The case is now unanswerable. *BMJ*. 1999;319(7214):865-–866. [PubMed: 10506022]

44. Dickersin K, Rennie D. The evolution of trial registries and their use to assess the clinical trial enterprise. *JAMA*. 2012;307(17):1861-–1864. [PubMed: 22550202]

45. 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. [PubMed: 21802902]

46. Gilbert R, Salanti G, Harden M, See S. Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. *Int J Epidemiol*. 2005;34(4):874-–887. [PubMed: 15843394]