# Tips for learners of evidence-based medicine: 4. Assessing heterogeneity of primary studies in systematic reviews and whether to combine their results

**Rose Hatala, Sheri Keitz, Peter Wyer, Gordon Guyatt, for the Evidence-Based Medicine Teaching Tips Working Group**

Clinicians wishing to quickly answer a clinical question may seek a systematic review, rather than searching for primary articles. Such a review is also called a meta-analysis when the investigators have used statistical techniques to combine results across studies. Databases useful for this purpose include the Cochrane Library (www.thecochranelibrary.com) and the ACP Journal Club (www.acpjc.org; use the search term "review"), both of which are available through personal or institutional subscription. Clinicians can use systematic reviews to guide clinical practice if they are able to understand and interpret the results.

Systematic reviews differ from traditional reviews in that they are usually confined to a single focused question, which serves as the basis for systematic searching, selection and critical evaluation of the relevant research.[1] Authors of systematic reviews use explicit methods to minimize bias and consider using statistical techniques to combine the results of individual studies. When appropriate, such pooling allows a more precise estimate of the magnitude of benefit or harm of a therapy. It may also increase the applicability of the result to a broader range of patient populations.

Clinicians encountering a meta-analysis frequently find the pooling process mysterious. Specifically, they wonder how authors decide whether the ranges of patients, interventions and outcomes are too broad to sensibly pool the results of the primary studies.

In this article we present an approach to evaluating potentially important differences in the results of individual studies being considered for a meta-analysis. These differences are frequently referred to as heterogeneity.[1] Our discussion focuses on the qualitative, rather than the statistical, assessment of heterogeneity (see Box 1).

Two concepts are commonly implied in the assessment of heterogeneity. The first is an assessment for heterogeneity within 4 key elements of the *design* of the original studies: the patients, interventions, outcomes and methods. This assessment bears on the question of whether pooling the results is at all sensible. The second concept relates to assessing heterogeneity among the *results* of the original studies. Even if the study designs are similar, the researchers must decide whether it is useful to combine the primary studies' results. Our discussion assumes a basic familiarity with how investigators present the magnitude[2,3] and precision[4] of treatment effects in individual randomized trials.

The tips in this article are adapted from approaches developed by educators with experience in teaching evidence-based medicine skills to clinicians.[1,5,6] A related article, intended for people who teach these concepts to clinicians, is available online at www.cmaj.ca/cgi/content/full/172/5/661/DC1.

## Clinician learners' objectives

### *Qualitative assessment of the design of primary studies*

- Understand the concepts of heterogeneity of study design among the individual studies included in a systematic review.

### *Qualitative assessment of the results of primary studies*

- Understand how to qualitatively determine the appropriateness of pooling estimates of effect from the individual studies by assessing (1) the degree of overlap of the confidence intervals around these point estimates of effect and (2) the disparity between the point estimates themselves.
- Understand how to estimate the "true" value of the estimate of effect from a graphic display of the results of individual studies.

**Teachers of evidence-based medicine:**

See the "Tips for teachers" version of this article online at www.cmaj.ca/cgi/content/full/172/5/661/DC1. It contains the exercises found in this article in fill-in-the-blank format, commentaries from the authors on the challenges they encounter when teaching these concepts to clinician learners and links to useful online resources.

DOI:10.1503/cmaj.1031920

**Box 1: Statistical assessments of heterogeneity**

Meta-analysts typically use 2 statistical approaches to evaluate the extent of variability in results between studies: Cochran's Q test and the $I^2$ statistic.

**Cochran's Q test**

- Cochran's Q test is the traditional test for heterogeneity. It begins with the null hypothesis that all of the apparent variability is due to chance. That is, the true underlying magnitude of effect (whether measured with a relative risk, an odds ratio or a risk difference) is the same across studies.
- The test then generates a probability, based on a $\chi^2$ distribution, that differences in results between studies as extreme as or more extreme than those observed could occur simply by chance.
- If the *p* value is low (say, less than 0.1) investigators should look hard for possible explanations of variability in results between studies (including differences in patients, interventions, measurement of outcomes and study design).
- As the *p* value gets very low (less than 0.01) we may be increasingly uncomfortable about using single best estimates of treatment effects.
- The traditional test for heterogeneity is limited, in that it may be underpowered (when studies have included few patients it may be difficult to reject the null hypothesis even if it is false) or overpowered (when sample sizes are very large, small and unimportant differences in magnitude of effect may nevertheless generate low *p* values).

**$I^2$ statistic**

- The $I^2$ statistic, the second approach to measuring heterogeneity, attempts to deal with potential underpowering or overpowering. $I^2$ provides an estimate of the percentage of variability in results across studies that is likely due to true differences in treatment effect, as opposed to chance.
- When $I^2$ is 0%, chance provides a satisfactory explanation for the variability we have observed, and we are more likely to be comfortable with a single pooled estimate of treatment effect.
- As $I^2$ increases, we get increasingly uncomfortable with a single pooled estimate, and the need to look for explanations of variability other than chance becomes more compelling.
- For example, one rule of thumb characterizes $I^2$ of less than 0.25 as low heterogeneity, 0.25 to 0.5 as moderate heterogeneity and over 0.5 as high heterogeneity.

## Tip 1: Qualitative assessment of the design of primary studies

Consider the following 3 hypothetical systematic reviews. For which of these systematic reviews does it make sense to combine the primary studies?

- A systematic review of all therapies for all types of cancer, intended to generate a single estimate of the impact of these therapies on mortality.
- A systematic review that examines the effect of different antibiotics, such as tetracyclines, penicillins and chloramphenicol, on improvement in peak expiratory flow rates and days of illness in patients with acute exacerbation of obstructive lung disease, including chronic bronchitis and emphysema.[7]
- A systematic review of the effectiveness of tissue plasminogen activator (tPA) compared with no treatment or placebo in reducing mortality among patients with acute myocardial infarction.[8]

Most clinicians would instinctively reject the first of these proposed reviews as overly broad but would be comfortable with the idea of combining the results of trials relevant to the third question. What about the second review? What aspects of the primary studies must be similar to justify combining their results in this systematic review?

Table 1 lists features that would be relevant to the question considered in the second review and categorizes them according to the 4 key elements of study design: the patients, interventions, outcomes and methods of the primary studies. Combining results is appropriate when the biology is such that across the range of patients, interventions, outcomes and study methods, one can anticipate more or less the same magnitude of treatment effect.

In other words, the judgement as to whether the primary studies are similar enough to be combined in a systematic review is based on whether the underlying pathophysiology would predict a similar treatment effect across the range of patients, interventions, outcomes and study methods of the primary studies. If you think back to the first systematic review — all therapies for all cancers — you probably recognize that there is significant variability in the

**Table 1: Relevant features of study design to be considered when deciding whether to pool studies in a systematic review (for a review examining the effect of antibiotics in patients with obstructive lung disease)**

| Patients | Interventions | Outcomes | Study methods |
|---|---|---|---|
| Patient age | Same antibiotic in all studies | Death | All randomized trials |
| Patient sex | Same class of antibiotic in all studies | Peak expiratory flow | Only blinded randomized trials |
| Type of lung disease (e.g., emphysema, chronic bronchitis) | Comparison of antibiotic with placebo | Forced expiratory volume in the first second | Cohort studies |
| | Comparison of one antibiotic with another | | |

pathophysiology of different cancers ("patients" in Table 1) and in the mechanisms of action of different cancer therapies ("interventions" in Table 1).

If you were inclined to reject pooling the results of the studies to be considered in the second systematic review, you might have reasoned that we would expect substantially different effects with different antibiotics, different infecting agents or different underlying lung pathology. If you were inclined to accept pooling of results in this review, you might argue that the antibiotics used in the different studies are all effective against the most common organisms underlying pulmonary exacerbations. You might also assert that the biology of an acute exacerbation of an obstructive lung disease (e.g., inflammation) is similar, despite variability in the underlying pathology. In other words, we would expect more or less the same effect across agents and across patients.

Finally, you probably accepted the validity of pooling results for the third systematic review — tPA for myocardial infarction — because you consider that the mechanism of myocardial infarction is relatively constant across a broad range of patients.

### *The bottom line*

- Similarity in the aspects of primary study design outlined in Table 1 (patients, interventions, outcomes, study methods) guides the decision as to whether it makes sense to combine the results of primary studies in a systematic review.
- The range of characteristics of the primary studies across which it is sensible to combine results is a matter of judgment based on the researcher's understanding of the underlying biology of the disease.

## Tip 2: Qualitative assessment of the results of primary studies

You should now understand that combining the results of different studies is sensible only when we expect more or less the same magnitude of treatment effects across the range of patients, interventions and outcomes that the investigators have included in their systematic review. However, even when we are confident of the similarity in design among the individual studies, we may still wonder whether the results of the studies should be pooled. The following graphic demonstration shows how to qualitatively assess the results of the primary studies to decide if meta-analysis (i.e., statistical pooling) is appropriate. You can find discussions of quantitative, or statistical, approaches to the assessment of heterogeneity elsewhere (see Box 1 or Higgins and associates[9]).

Consider the results of the studies in 2 hypothetical systematic reviews (Fig. 1A and Fig. 1B). The central vertical line, labelled "no difference," represents a treatment effect of 0. This would be equivalent to a risk ratio or relative risk of 1 or an absolute or relative risk reduction of 0.[2] Values to the left of the "no difference" line indicate that the treatment is superior to the control, whereas those to the right of the line indicate that the control is superior to the treatment. For each of the 4 studies represented in the figures, the dot represents the point estimate of the treatment effect (the value observed in the study), and the horizontal line represents the confidence interval around that observed effect. For which systematic review does it make sense to combine results? Decide on the answer to this question before you read on.

You have probably concluded that pooling is appropriate

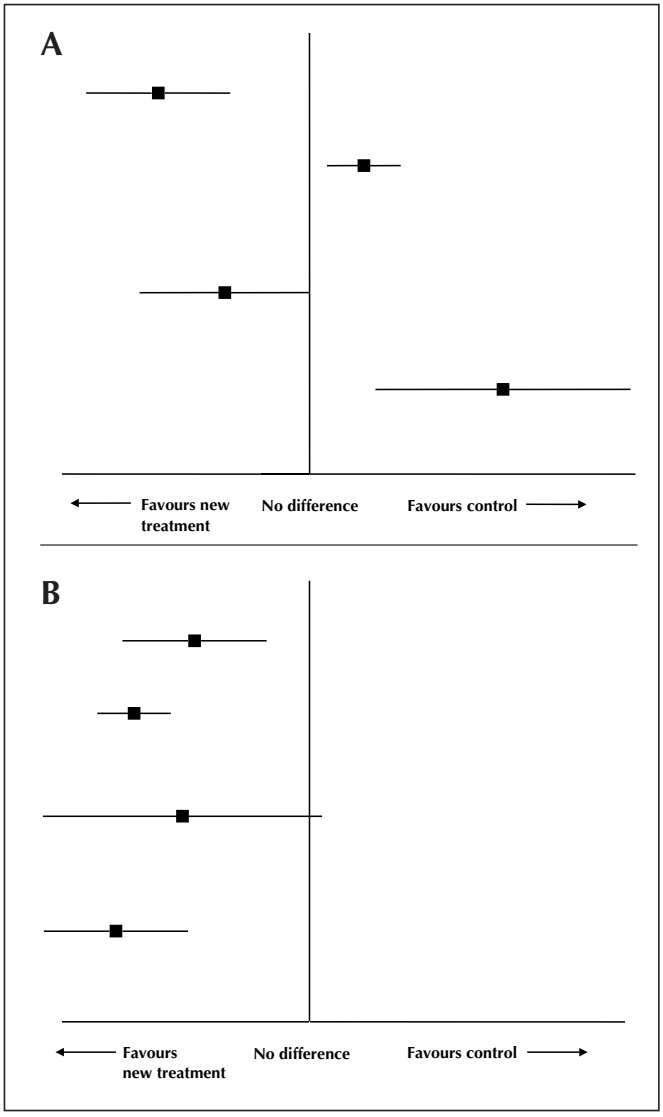


**Fig. 1: Results of the studies in 2 hypothetical systematic reviews.** The central vertical line represents a treatment effect of 0. Values to the left of this line indicate that the treatment is superior to the control, whereas those to the right of the line indicate that the control is superior to the treatment. For each of the 4 studies in each figure, the dot represents the point estimate of the treatment effect (the value observed in the study), and the horizontal line represents the confidence interval around that observed effect.

for the studies represented in Fig. 1B but not for those represented in Fig. 1A. Can you explain why? Is it because the point estimates for the studies in Fig. 1A lie on opposite sides of the "no difference" line, whereas those for the studies in Fig. 1B lie on the same side of the "no difference" line?

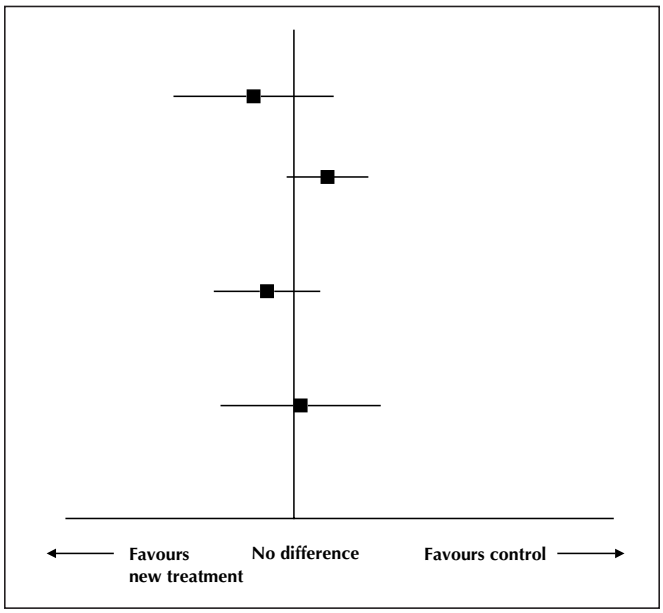


**Fig. 2: Point estimates and confidence intervals for 4 studies.** Two of the point estimates favour the new treatment, and the other 2 point estimates favour the control. Investigators doing a systematic review with these 4 studies would be satisfied that it is appropriate to pool the results.

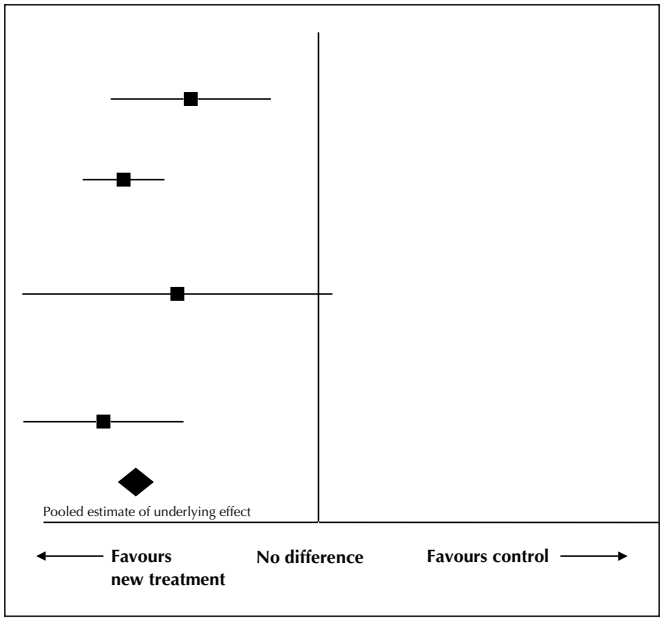


**Fig. 3: Results of the hypothetical systematic review presented in Fig. 1B.** The pooled estimate at the bottom of the chart (large diamond) provides the best guess as to the underlying treatment effect. It is centred on the midpoint of the area of overlap of the confidence intervals around the estimates of the individual trials.

Before you answer this question, consider the studies represented in Fig. 2. Here, the point estimates of 2 studies are on the "favours new treatment" side of the "no difference" line, and the point estimates of 2 other studies are on the "favours control" side. However, all 4 point estimates are very close to the "no difference" line, and, in this case, investigators doing a systematic review will be satisfied that it is appropriate to pool the results. Therefore, it is not the position of the point estimates relative to the "no difference" line that determines the appropriateness of pooling.

There are 2 criteria for not combining the results of studies in a meta-analysis: highly disparate point estimates and confidence intervals with little overlap, both of which are exemplified by Fig. 1A. When pooling *is* appropriate on the basis of these criteria, where is the best estimate of the underlying magnitude of effect likely to be? Look again at Fig. 1B and make a guess. Now look at Fig. 3.

The pooled estimate at the bottom of Fig. 3 is centred on the midpoint of the area of overlap of the confidence intervals around the estimates of the individual trials. It provides our best guess as to the underlying treatment effect. Of course, we cannot actually know the "truth" and must be content with potentially misleading estimates. The intent of a meta-analysis is to include enough studies to narrow the confidence interval around the resulting pooled estimate sufficiently to provide estimates of benefit for our patients in which we can be confident. Thus, our best estimate of the truth will lie in the area of overlap among the confidence intervals around the point estimates of treatment effect presented in the primary studies.

What is the clinician to do when presented with results such as those in Fig. 1A? If the investigators have done a good job of planning and executing the meta-analysis, they will provide some assistance.[6] Before examining the study results in detail, they will have generated a priori hypotheses to explain the heterogeneity in magnitude of effect across studies that they are liable to encounter. These hypotheses will include differences in patients (effects may be larger in sicker patients), in interventions (larger doses may result in larger effects), in outcomes (longer follow-up may diminish the magnitude of effect) and in study design (methodologically weaker studies may generate larger effects).

The investigators will then have examined the extent to which these hypotheses can explain the differences in magnitude of effect across studies. These subgroup analyses may be misleading, but if they meet 7 criteria suggested elsewhere[10] (see Box 2), they may provide credible and satisfying explanations for the variability in results.

## *The bottom line*

- Readers can decide for themselves whether there is clinically important heterogeneity among the results of primary studies through a qualitative assessment of the graphic results. This assessment is based on the amount

**Box 2: Questions to ask when evaluating a subgroup analysis in a meta-analysis[10]**

- Was the subgroup comparison based on a within-study, rather than a between-study, comparison?
- Is the magnitude of the difference in effect between subgroups large?
- Is the effect consistent across studies?
- Is the difference in effect statistically significant?
- Was the subgroup analysis planned in advance by the trialists?
- Were many subgroup analyses performed and selectively reported?
- Is the difference in effect in the subgroup supported by a biological hypothesis?

of disparity among the individual point estimates and the degree of overlap among the confidence intervals.

## Conclusions

Understanding the concept of heterogeneity in a systematic review or meta-analysis is central to a full appreciation of the implications of such reviews for clinical practice. We have presented 2 tips aimed at helping clinical readers overcome commonly encountered difficulties in understanding this concept.

This article has been peer reviewed.

From the Department of Medicine, University of British Columbia, Vancouver, BC (Hatala); Durham Veterans Affairs Medical Center and Duke University Medical Center, Durham, NC (Keitz); the Columbia University College of Physicians and Surgeons, New York, NY (Wyer); and the Departments of Medicine and of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont. (Guyatt)

*Competing interests:* None declared.

*Contributors:* Rose Hatala modified the original ideas for tips 1 and 2, drafted the manuscript, coordinated input from reviewers and field-testing, and revised all drafts. Sheri Keitz used all of the tips as part of a live teaching exercise and submitted comments, suggestions and the possible variations that are described in the article. Peter Wyer reviewed and revised the final draft of the manuscript to achieve uniform adherence with format specifications. Gordon Guyatt developed the original ideas for tips 1 and 2, reviewed the manuscript at all phases of development, contributed to the writing as a coauthor, and, as general editor, reviewed and revised the final draft of the manuscript to achieve accuracy and consistency of content.

## References

1. Oxman A, Guyatt G, Cook D, Montori V. Summarizing the evidence. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 155-73.
2. Barratt A, Wyer PC, Hatala R, McGinn T, Dans AL, Keitz S, et al, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. *CMAJ* 2004;171(4):353-8.
3. Guyatt G, Cook D, Devereaux PJ, Meade M, Straus S. Therapy. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 55-79.
4. Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Moyer V, et al, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for learners of evidence-based medicine: 2. Measures of precision (confidence intervals). *CMAJ* 2004;171(6):611-5.
5. Wyer PC, Keitz S, Hatala R, Hayward R, Barratt A, Montori V, et al. Tips for learning and teaching evidence-based medicine: introduction to the series. *CMAJ* 2004;171(4):347-8.
6. Montori V, Hatala R, Guyatt G. Summarizing the evidence: evaluating differences in study results. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 547-52.
7. Saint S, Bent S, Vittinghoff E, Grady D. Antibiotics in chronic obstructive pulmonary disease exacerbations. *JAMA* 1995;273:957-60.
8. Held PH, Teo KK, Yusuf S. Effects of tissue-type plasminogen activator and anisoylated plasminogen streptokinase activator complex on mortality in acute myocardial infarction. *Circulation* 1990;82:1668-74.
9. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
10. Oxman A, Guyatt G. When to believe a subgroup analysis. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 553-65.

***Correspondence to:*** *Dr. Peter C. Wyer, 446 Pelhamdale Ave., Pelham NY 10804; fax 914 738-9368; pwyer@att.net*

***Members of the Evidence-Based Medicine Teaching Tips Working Group:*** *Peter C. Wyer (project director), College of Physicians and Surgeons, Columbia University, New York, NY; Deborah Cook, Gordon Guyatt (general editor), Ted Haines, Roman Jaeschke, McMaster University, Hamilton, Ont.; Rose Hatala (internal review coordinator), University of British Columbia, Vancouver, BC; Robert Hayward (editor, online version), Bruce Fisher, University of Alberta, Edmonton, Alta.; Sheri Keitz (field test coordinator), Durham Veterans Affairs Medical Center and Duke University Medical Center, Durham, NC; Alexandra Barratt, University of Sydney, Sydney, Australia; Pamela Charney, Albert Einstein College of Medicine, Bronx, NY; Antonio L. Dans, University of the Philippines College of Medicine, Manila, The Philippines; Barnet Eskin, Morristown Memorial Hospital, Morristown, NJ; Jennifer Kleinbart, Emory University School of Medicine, Atlanta, Ga.; Hui Lee, formerly Group Health Centre, Sault Ste. Marie, Ont. (deceased); Rosanne Leipzig, Thomas McGinn, Mount Sinai Medical Center, New York, NY; Victor M. Montori, Mayo Clinic College of Medicine, Rochester, Minn.; Virginia Moyer, University of Texas, Houston, Tex.; Thomas B. Newman, University of California, San Francisco, San Francisco, Calif.; Jim Nishikawa, University of Ottawa, Ottawa, Ont.; Kameshwar Prasad, Arabian Gulf University, Manama, Bahrain; W. Scott Richardson, Wright State University, Dayton, Ohio; Mark C. Wilson, University of Iowa, Iowa City, Iowa*

**Articles to date in this series**