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

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

Clinicians wishing to quickly answer a clinical question may look for 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 quickly access many such reviews and can use them to guide clinical practice — if they are able to understand and interpret the results.

Educators may find opportunities to evaluate such articles with a group of learners in many teaching contexts, including morning reports, journal clubs or ward rounds. During a discussion of almost any systematic review the following question will arise: "Would it make sense to pool the results of the primary studies in a meta-analysis?"

In this article we present an approach to helping clinicians understand the issue of heterogeneity of individual study results in systematic reviews. As with other articles in this series, clinical educators experienced in teaching evidence-based medicine developed the tips and have used them extensively. A full description of the development of the tips presented in this series, as well as pertinent background information, has been presented elsewhere.[1]

For each of the 2 tips in this article, we have provided guidance on when to use the tip, the teaching script for the tip, a "bottom line" section and a summary card. For each tip we have identified the appropriate level of learner experience and provided estimates of the time required for the exercise.

This article addresses 2 stumbling blocks to understanding a systematic review that are commonly encountered by learners. Both revolve around the learner's need to identify and evaluate potentially important differences in the results of individual studies being combined in a meta-analysis, frequently termed heterogeneity.[2] The tips are geared to learners who are familiar with how investigators should present the magnitude[3–5] and precision[6,7] of treatment effects in a study with binary outcomes (the risk ratio or relative risk reduction) and who have previously assessed a few therapy articles.

That "heterogeneity" implies 2 distinct concepts constitutes the first stumbling block in understanding the term. The first concept is that of heterogeneity among the patients, interventions, outcomes and methodologies of the original studies. This bears on whether any sort of pooling is at all sensible. We refer to these 4 key elements as components of the "study design." The second concept is that of heterogeneity in the magnitude of effect across studies. The extent of variability in magnitude of effect bears on whether — assuming that combining the results seems reasonable from the standpoint of study design — the results of the individual studies raise serious questions about the appropriateness of pooling.

The second stumbling block arises in the process of evaluating this latter question about the degree of variability in magnitude of effect between the primary studies. Many learners are intimidated by meta-analysis and, as a result, tend to fixate on the most intimidating aspect, the method of combining results from different studies. The apparent mystery of statistical tests for heterogeneity tends to reinforce the clinician learner's notion that "This is something for experts, not for me." The second teaching tip attempts to overcome this "numero-phobia" by suggesting a quick and easy visual assessment of the results.

**Other available resources**

- A companion version of this article directed to learners of evidence-based medicine has been published in *CMAJ* and is available online through *eCMAJ* (www.cmaj.ca/cgi/content/full/172/5/661/DC1)
- An interactive version of this article, as well as other tools and resources, is available at www.ebmtips.net/heterogeneity001.asp.

## Teaching tip 1: Qualitative assessment of the design of primary studies

### *When to use this tip*

This tip focuses on whether it makes conceptual sense to have undertaken the systematic review. It is geared to learners who are encountering a systematic review for the first time, but it may also be useful for any learner who has not fully grasped the concept of heterogeneity. It takes 10 to 15 minutes to complete. This tip is intended to help learners meet the following specific objective:

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

The tip emphasizes the importance of assessing heterogeneity in the designs of the primary studies. With more advanced learners, the instructor can proceed directly to teaching tip 2, which incorporates the concepts of this first tip.

### *The script*

Have the learners consider 3 hypothetical systematic reviews. Choose the examples such that for one review it is unconscionable to combine the primary studies, for the second it is debatable and for the third it is eminently reasonable. The following examples fulfill these criteria.

- 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.[8] This is an attractive example because, although most learners accept the question as appropriate, there is sufficient breadth among the treatments and patients that it is debatable whether it makes sense to combine the primary studies.
- 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.[9]

Ask the learners, "For which of these systematic reviews does it make sense to combine the primary studies?" The learners will unanimously reject the first proposed review. Many will also reject the second review, although with a little less certainty, and there may be some dissent. Most learners are comfortable with the third proposed review.

Guide the group toward understanding the basis on which they made their decisions. The typical answer will be some variant of a statement that the first 2 reviews are "too broad" or that certain aspects of the designs of the primary studies in the first 2 reviews are "not similar enough" to combine them. The following exercise will help lead the group toward the correct answer, that combining results is appropriate only when the biology is such that, across the range of patients, interventions, outcomes and study methodologies, one can anticipate more or less the same magnitude of treatment effect. Using the second example, the systematic review in which combining the primary studies is debatable, ask the learners what aspects of the primary studies must be similar if their results are to be combined in a systematic review. Use a blackboard to organize the answers into a table with 4 columns, one for each of the 4 key elements of study design: patients, interventions, outcomes and methodologies (Table 1A). Write in these column headings after the group has given their answers. The completed table, including characteristic learner responses, is shown as Table 1B.

It should now be apparent that 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 methodologies of the primary studies. The group is led to appreciate that they rejected the first systematic review — all therapies for all cancers — because of the variability in the pathophysiology of different cancers ("patients" in Table 1B) and in the mechanisms of action of different cancer therapies ("interventions" in Table 1B).

Similarly, you lead the group to understand that the implied logic of those who rejected the second systematic review is that we might expect substantially different effects with different antibiotics, different infecting agents or different underlying lung pathology. Those who were ready to accept the suitability of pooling for the second systematic review will have reasoned differently. They might argue that the antibiotics used in the different studies are all effective against the most common organisms underlying pulmonary exacerbations. They 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. Thus, they would argue, we would expect more or less the same effect across agents and across patients, and combining results is therefore appropriate.

Finally, the group is led to appreciate that they unanimously accepted the third proposed systematic review — tPA for myocardial infarction — because the mechanism of myocardial infarction is relatively consistent across a broad range of patients.

### *The bottom line*

- Similarity across patients, interventions, outcomes and methodologies guides the assessment of appropriateness of combining the results of primary studies in a systematic review.
- Whether studies are similar enough depends on whether we would anticipate more or less the same magnitude of treatment effect across studies.
- The range of characteristics of the primary studies across which it is sensible to combine results is a matter of judgement based on the researcher's understanding of the underlying biology of the disease.

See Appendix 1 for the summary card for this tip.

### *Extension for advanced learners*

Learners who have encountered meta-analysis previously will grasp that looking directly at the results of the individual studies also indicates whether the systematic review is sensible (see also teaching tip 2, below). If significant heterogeneity exists among the results of the primary studies, there may be important differences in the individual study designs. If there is minimal heterogeneity, it should be appropriate to combine the results.

You might then ask the group, "What is the advantage of a systematic review? Why do we read them?" The learners will bring up numerous reasons, often focusing on the larger sample size. You can then lead the group toward a second, equally important advantage of a systematic review: the wide applicability of the results to clinical practice.

Ask the learners to think of the last patient they saw with the disorder covered by the systematic review. Choose one primary study from the systematic review and ask each learner if his or her patient would have qualified for inclusion in that primary study. Does the patient qualify for inclusion in any of the primary studies (i.e., would the patient be included in the meta-analysis)? Once the learners grasp that the results from the systematic review are more applicable to their patients than are the results from an individual study (because the systematic review embraces variability across study patients, interventions and outcomes), they will have added another criterion to their definition of "similar enough" for the assessment of heterogeneity among study designs.

## Teaching tip 2: Qualitative assessment of the results of primary studies

### *When to use this tip*

This tip hinges on the learners' original question: "How do we know if it is sensible to combine the results of the primary studies?" It is appropriate to undertake this tip once they have grasped the content of teaching tip 1. The tip is intended to help learners meet the following specific objectives:

- Understand how to qualitatively determine the appropriateness of pooling estimates of effect from the individual studies by assessing (a) the degree of overlap of the confidence intervals around these point estimates of effect and (b) the disparity between the point estimates themselves.

**Table 1A: Relevant features of study design to be considered when deciding whether to pool studies in a systematic review**

| Patients | Interventions | Outcomes | Study methodologies |
|---|---|---|---|

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

| Patients | Interventions | Outcomes | Study methodologies |
|---|---|---|---|
| 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 | | |

- Understand how to estimate the "true" value of the pooled estimate of effect from a graphic display of the results of individual studies.

The learners discover in the course of tip 1 that combining the results of different studies is sensible only when we expect similar results across the range of patients, interventions, outcomes and methodologies that the investigators have included in their systematic review. We are seldom completely confident about the similarity among the designs of individual studies, however, and thus may still wonder, "Should the results of the studies be pooled?" We can close the loop by asking, "Were the data consistent with the original assumption that the results would be more or less the same across studies?" 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. It should strongly appeal to visual learners. The tip will take 15 to 30 minutes to complete, and the presentation may be facilitated by drawing the graphics on an overhead or blackboard before the teaching session begins.

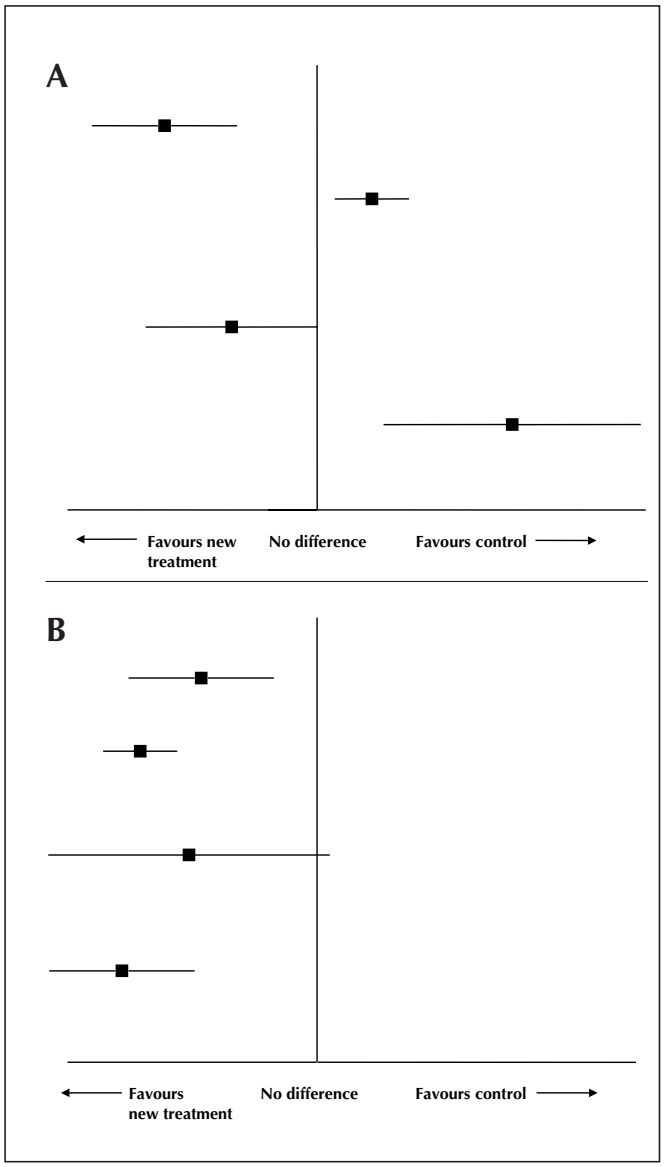


**Fig. 1: Results of the studies in 2 hypothetical systematic reviews.** The central vertical line represents a treatment effect of 0. Values to the left of this line indicate that the treatment is superior to the control, whereas those to the right of the line indicate that the control is superior to 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.

### *The script*

Draw 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 (a risk ratio of 1). 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. Ensure that the learners understand that the dots represent the point estimates (the observed values of the treatment effect) and the lines represent the confidence intervals around those observed effects.

Ask the group, "For which systematic review does it make sense to combine results?" The group gets this right pretty quickly, and you then inquire "Why?"

Some learners will point to the fact that 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. Most times, this suggestion will meet with the group's approval. In response, draw the results of another hypothetical review in which the point estimates of 2 studies are on the "favours new treatment" side of the "no difference" line, the point estimates of 2 other studies are on the "favours control" side, with all 4 point estimates very close to the "no difference" line (Fig. 2).

Since the group is happy combining these results, the proposed criterion — that we reject combining if the results are on different sides of the "no difference" line — fails. The group eventually arrives at the 2 criteria for rejecting combining: highly disparate point estimates and confidence intervals with little overlap.

Once the learners have struggled to articulate "Why," ask one of them to draw, at the top of Fig. 1B, the results that would be obtained if all eligible patients in the world were included in a single, unbiased study and if the "truth" is that there is benefit to the new treatment. Eventually the graph will look like Fig. 3, with the hypothetical "truth" for the results lying at the midpoint of the overlapping confidence intervals.

Conclude this tip by emphasizing to the learners that the "truth" box at the top of Fig. 3 is a theoretical represen-

tation of what we are striving for but can never be sure of attaining. Since we can never be sure of attaining the "truth" and must be content with potentially misleading estimates, the pooled estimate that is commonly reported, as shown at the bottom of Fig. 3, represents our best guess as to the underlying treatment effect, obtained by combining the results of primary studies in a meta-analysis.

The intent of a meta-analysis is to include enough studies to narrow the confidence interval sufficiently to provide estimates of benefit for our patients in which we can be confident. Our best estimate will lie in the area of overlap among the confidence intervals around the point estimates of the primary studies. When the confidence intervals do not overlap, as in Fig. 1A, it does not make sense to pool the results of the primary studies, as there is no common area of potential "truth."

## The bottom line

- A simple plot of the results of studies to be considered for combining in a systematic review allows the learners to estimate the appropriateness of such pooling.
- The amount of disparity between the individual point estimates and the degree of overlap among the confidence intervals around the point estimates determine whether the results should be combined.
- The midpoint of the overlapping confidence intervals is likely to be close to the estimate of effect obtained by pooling.

See Appendix 1 for the summary card for this tip.

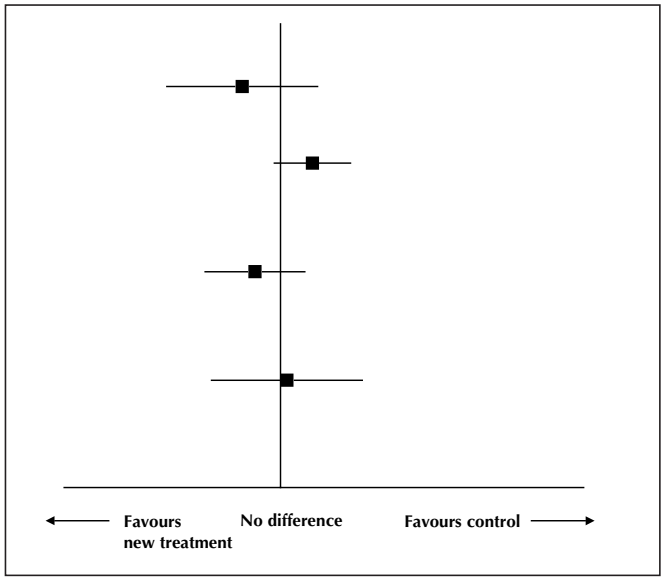


**Fig. 2: Point estimates and confidence intervals for 4 studies.** Two of the point estimates favour the new treatment, and the other 2 point estimates favour the control. Investigators doing a systematic review with these 4 studies would be satisfied that it is appropriate to pool the results.

## Extensions for advanced learners

Some learners will wonder why, despite apparent similarities in the designs of the primary studies (patients, interventions, outcomes, methodologies), the results may be dissimilar. Allow the group to hypothesize some causes for these differences. Then, use Fig. 2 to graphically highlight that if any of these hypotheses are correct, the primary study results may fall into separable groups. One advantage of a meta-analysis over a primary study is the ability to define why study results vary (i.e., to explain the observed heterogeneity and hence to better understand the primary disorder and its treatment).

The statistical test for heterogeneity is a frequent source of confusion for clinicians and learners. A prerequisite for understanding this concept is some knowledge of the significance of *p* values.[10] In the assessment of heterogeneity, the null hypothesis is that there is no difference in the magnitude of the treatment effect across the primary studies. Expressed in terms of the risk ratio (RR), this would mean that $RR_1 = RR_2 = RR_3$ and so on. Given this definition, ask the learners what *p* value should be assigned to the results of the studies represented in Fig. 1A. Someone will suggest that it should be less than 0.05, and someone else may correctly suggest that it is extremely small. Similarly, the learners catch on that the comparable *p* value for the results of the studies represented in Fig. 1B

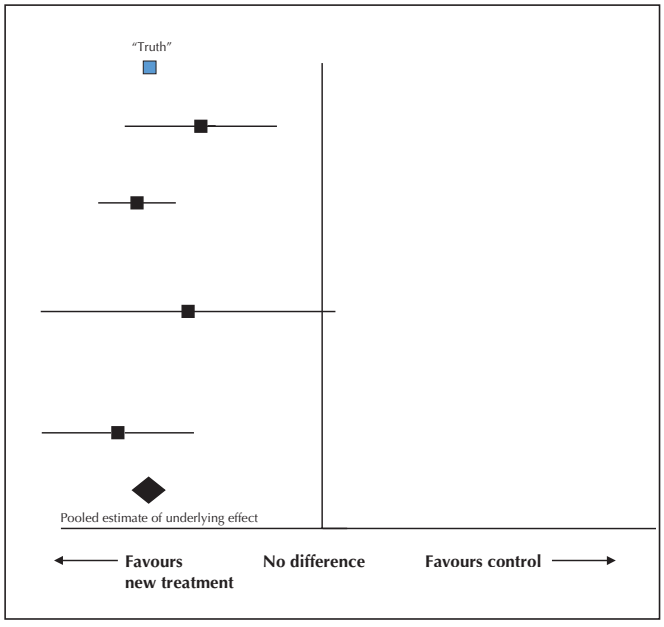


**Fig. 3: Results of the hypothetical systematic review presented in Fig. 1B.** The pooled estimate at the bottom of the chart (large diamond) provides the best guess as to the underlying treatment effect. It is centred on the midpoint of the area of overlap of the confidence intervals around the estimates of the individual trials. The "truth" box at the top of the figure is a theoretical representation of the information researchers are seeking but can never be sure of attaining.

is greater than 0.05, because they are consistent with the null hypothesis representing the underlying truth. Some will realize that the *p* value for these results is very close to 1.0.

In general, when teaching clinical learners, we avoid detailed discussions of quantitative, or statistical, approaches to the assessment of heterogeneity. Sometimes, however, we draw on summaries such as that provided in Box 1 and by Higgins and associates.[11]

**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.

## Report on field-testing

One of us (S.K.), an experienced teacher of evidence-based medicine, field-tested these tips in November 2001 with 12 medical interns and residents during a 1.5-hour teaching session. Eight participants were very naive learners with little prior exposure to the principles of evidence-based medicine, and 4 had moderate prior exposure to these principles.

Both tips worked entirely as intended. The study examples in tip 1 and the graphic examples in tip 2 effectively clarified the principles of heterogeneity. Through the use of examples that relied on common sense, rather than formulas and statistical models, the learners were able to bypass their fears about meta-analyses.

Before using either tip, however, the learners required a discussion of the definitions and differences between a narrative systematic review (qualitative) and a meta-analysis (quantitative). Even the more advanced learners were not crystal clear on this distinction. Without clarity on the learners' part regarding these definitions, even an experienced teacher would not be able to teach the tips, as the tips are essentially about the interface between a narrative systematic review and a meta-analysis.

S.K. extended the teaching session by including an assessment of 2 articles after teaching tips 1 and 2. The residents were divided into 2 groups, each responsible for discussing one of the articles and then describing whether the set of primary studies could or should be combined in a meta-analysis. This exercise led to an "Aha!" moment in the teaching session. The residents were amazed that they could now understand at a glance the graphic display of the primary studies' results.

When asked about the relevance of the content and clarity of the presentations, the learners assigned mean scores of 8.5 and 9.0 to tips 1 and 2 respectively using a visual analogue scale ranging from 0 to 10. The residents felt that the session had improved their understanding of the different types of reviews, although they highlighted the importance of modifying the teaching session to first cover this background information. The key outcome for the residents was that the teaching session diminished their fear and increased their confidence in their ability to interpret a meta-analysis.

## Conclusions

Understanding the concept of heterogeneity in a systematic review or meta-analysis is central to finding these articles relevant to clinical practice. We have presented 2 teaching tips, developed and used by experienced clinician-educators, that help overcome the learner difficulties commonly encountered in teaching 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 reported 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. 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.
2. 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.
3. 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 teachers of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. *CMAJ* 2004;171(4):Online-1 to Online 8. Available: www.cmaj.ca/cgi/content/full/171/4/353/DC1 (accessed 2005 Feb 16).
4. 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(1):353-8.
5. 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 of evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 55-79.
6. Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Guyatt G, et al for the Evidence-Based Medicine Teaching Tips Working Group. Tips for teachers of evidence based medicine: 2. Confidence intervals and *p* values. *CMAJ* 2004;171(6):Online 1 to Online 12. Available: www.cmaj.ca/cgi/content/full/171/6/611/DC1 (accessed 2005 Feb 16).
7. 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.
8. Saint S, Bent S, Vittinghoff E, Grady D. Antibiotics in chronic obstructive pulmonary disease exacerbations. *JAMA* 1995;273:957-60.
9. 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.
10. Guyatt G, Jaeschke R, Cook D, Walter S. Therapy and understanding the results: hypothesis testing. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual of evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 329-38.
11. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.

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



**Articles to date in this series**

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 teachers of evidence-based medicine: 1. Relative risk reduction, absolute risk reduction and number needed to treat. Available: www.cmaj.ca/cgi/content/full/171/4/353/DC1.

Montori VM, Kleinbart J, Newman TB, Keitz S, Wyer PC, Guyatt G, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for teachers of evidence-based medicine: 2. Confidence intervals and *p* values. Available: www.cmaj.ca/cgi/content/full/171/6/611/DC1.

McGinn T, Wyer PC, Newman TB, Keitz S, Leipzig R, Guyatt G, for the Evidence-Based Medicine Teaching Tips Working Group. Tips for teachers of evidence-based medicine: 3. Understanding and calculating kappa. Available: www.cmaj.ca/cgi/content/full/171/11/1369/DC1.

## Appendix 1: Summary cards for 2 teaching tips on heterogeneity

This appendix has been designed so that it can be printed on one 8½ × 11 inch page. The individual summary cards can then be cut out, if desired, for use during teaching sessions.

### Teaching tip 1: Qualitative assessment of the design of primary studies

**Scenario:** Consider a hypothetical systematic review for which combining the studies is implausible. Then propose 2 additional systematic reviews in which such a combining is debatable and eminently reasonable, respectively. The learners must rate the appropriateness of combining the results for each of the successive reviews.

1. Describe the first hypothetical review as addressing the effect on mortality of all therapies for all types of cancer.
2. Describe the second review as addressing the effect of antibiotic therapy on peak expiratory flow rate and duration of illness in patients with acute exacerbations of obstructive lung disease.
3. Describe the third review as addressing the effect on mortality of tissue plasminogen activator in patients with acute myocardial infarction.
4. For the second hypothetical review, have the learners identify the characteristics of the studies that need to be similar for combining their results to be appropriate. Write their responses in the appropriate columns of a table with headings "Patients," "Interventions," "Outcomes" and "Study methodologies."

**Summary points**

- Similarity across patients, interventions, outcomes and methodologies guides the assessment of the appropriateness of combining the results of the studies included in a systematic review.
- Whether studies are similar enough depends on whether one would expect the same magnitude of treatment effect across the studies.
- The range of characteristics of the primary studies across which it is sensible to combine results is a matter of judgement based on the researcher's understanding of the biology of the disease.

### Teaching tip 2: Qualitative assessment of the results of primary studies

**Scenario:** Consider the results of the individual studies to be included in a series of systematic reviews. In some, the direction and magnitude of effects are similar and the confidence intervals overlap, whereas in others the effects are importantly different, and the confidence intervals overlap minimally, if at all. The learners must judge the appropriateness of combining the studies by looking at a plot of the observed effects and the confidence intervals in a figure.

1. Construct 2 meta-view plots representing 2 sets of 4 studies each. In one plot, the study results vary widely in size and direction of effect, and the confidence intervals do not overlap. In the other plot, the study results are all in the same direction, and the confidence intervals do overlap. The learners must decide for which review they would combine the results, giving reasons for their answers.
2. Provide a third example in which the results fall on different sides of the "no difference" line but are otherwise close together, with substantial overlap of the confidence intervals. The learners discover that it is the absolute difference in magnitude of results, not whether they lie in the same direction, that (together with the extent of overlap of the confidence intervals) determines whether they should be combined.
3. Have the learners estimate the true (i.e., pooled) result of the meta-analysis in an example in which the results meet both of the above criteria. They learn to choose the midpoint of the overlap of the different confidence intervals.

**Summary points**

- A simple plot of the results of studies to be considered for combining in a systematic review allows the learners to estimate the appropriateness of such pooling.
- The amount of disparity among the individual point estimates and the degree of overlap among the confidence intervals around the point estimates determine whether the results should be combined.
- The midpoint of the overlapping confidence intervals is likely to be close to the estimate of the effect obtained by pooling.