Tips for learners of evidence-based medicine: 2. Measures of precision (confidence intervals)

Victor M. Montori, Jennifer Kleinbart, Thomas B. Newman, Sheri Keitz, Peter C. Wyer, Virginia Moyer, Gordon Guyatt, for the Evidence-Based Medicine Teaching Tips Working Group

In the first article in this series,[1] we presented an approach to understanding how to estimate a treatment's effectiveness that covered relative risk reduction, absolute risk reduction and number needed to treat. But how precise are these estimates of treatment effect?

In reading the results of clinical trials, clinicians often come across 2 related but different statistical measures of an estimate's precision: *p* values and confidence intervals. The *p* value describes how often apparent differences in treatment effect that are as large as or larger than those observed in a particular trial will occur in a long run of identical trials if in fact no true effect exists. If the observed differences are sufficiently unlikely to occur by chance alone, investigators reject the hypothesis that there is no effect. For example, consider a randomized trial comparing diuretics with placebo that finds a 25% relative risk reduction for stroke with a *p* value of 0.04. This *p* value means that, if diuretics were in fact no different in effectiveness than placebo, we would expect, by the play of chance alone, to observe a reduction — or increase — in relative risk of 25% or more in 4 out of 100 identical trials.

Although they are useful for investigators planning how large a study needs to be to demonstrate a particular magnitude of effect, *p* values fail to provide clinicians and patients with the information they most need, i.e., the range of values within which the true effect is likely to reside. However, confidence intervals provide exactly that information in a form that pertains directly to the process of deciding whether to administer a therapy to patients. If the range of possible true effects encompassed by the confidence interval is overly wide, the clinician may choose to administer the therapy only selectively or not at all.

Confidence intervals are therefore the topic of this article. For a nontechnical explanation of *p* values and their limitations, we refer interested readers to the *Users' Guides to the Medical Literature*.[2]

As with the first article in this series,[1] we present the information as a series of "tips" or exercises. This means that you, the reader, will have to do some work in the course of reading the article. The tips we present here have been adapted from approaches developed by educators experienced in teaching evidence-based medicine skills to clinicians.[2-4] A related article, intended for people who teach these concepts to clinicians, is available online at www.cmaj.ca/cgi/content/full/171/6/611/DC1.

Clinician learners' objectives

Making confidence intervals intuitive

- Understand the dynamic relation between confidence intervals and sample size.

Interpreting confidence intervals

- Understand how the confidence intervals around estimates of treatment effect can affect therapeutic decisions.

Estimating confidence intervals for extreme proportions

- Learn a shortcut for estimating the upper limit of the 95% confidence intervals for proportions with very small numerators and for proportions with numerators very close to the corresponding denominators.

Tip 1: Making confidence intervals intuitive

Imagine a hypothetical series of 5 trials (of equal duration but different sample sizes) in which investigators have experimented with treatments for patients who have a particular condition (elevated low-density lipoprotein cholesterol) to determine whether a drug (a novel cholesterol-lowering agent) would work better than a placebo to prevent strokes (Table 1A). The smallest trial enrolled only

Teachers of evidence-based medicine:

See the "Tips for teachers" version of this article online at www.cmaj.ca/cgi/content/full/171/6/611/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.1031667

8 patients, and the largest enrolled 2000 patients, and half of the patients in each trial underwent the experimental treatment. Now imagine that all of the trials showed a relative risk reduction for the treatment group of 50% (meaning that patients in the drug treatment group were only half as likely as those in the placebo group to have a stroke). In each individual trial, how confident can we be that the true value of the relative risk reduction is important for patients (i.e., "patient-important")?[5] If you were to look at the studies individually, which ones would lead you to recommend the treatment unequivocally to your patients?

Most clinicians might intuitively guess that we could be more confident in the results of the larger trials. Why is this? In the absence of bias or systematic error, the results of a trial can be interpreted as an estimate of the true magnitude of effect that would occur if all possible eligible patients had been included. When only a few of these patients are included, the play of chance alone may lead to a result that is quite different from the true value. Confidence intervals are a numeric measure of the range within which such variation is likely to occur. The 95% confidence intervals that we often see in biomedical publications represent the range within which we are likely to find the underlying true treatment effect.

To gain a better appreciation of confidence intervals, go back to Table 1A (don't look yet at Table 1B!) and take a guess at what you think the confidence intervals might be for the 5 trials presented. In a moment you'll see how your estimates compare to 95% confidence intervals calculated using a formula, but for now, try figuring out intervals that you intuitively feel to be appropriate.

Now, consider the first trial, in which 2 out of 4 patients who receive the control intervention and 1 out of 4 patients who receive the experimental treatment suffer a stroke. The risk in the treatment group is half that in the control group, which gives us a relative risk of 50% and a relative risk reduction of 50% (see Table 1A).[1,6]

Given the substantial relative risk reduction, would you be ready to recommend this treatment to a patient? Before you answer this question, consider whether it is plausible, with so few patients in the study, that the investigators might just have gotten lucky and the true treatment effect is really a 50% *increase* in relative risk. In other words, is it plausible that the true event rate in the group that received treatment was 3 out of 4 instead of 1 out of 4? If you accept that this large, harmful effect might represent the underlying truth, would you also accept that a relative risk reduction of 90%, i.e., a very large benefit of treatment, is consistent with the experimental data in these few patients? To the extent that these suggestions are plausible, we can intuitively create a range of plausible truth of "-50% to 90%" surrounding the relative risk reduction of 50% that was actually observed.

Now, do this for each of the other 4 trials. In the trial with 20 patients in each group, 10 of those in the control group suffered a stroke, as did 5 of those in the treatment group. Both the relative risk and the relative risk reduction are again 50%. Do you still consider it plausible that the true event rate in the treatment group is 15 out of 20 rather than 5 out of 20 (the same proportions as we considered in the smaller trial)? If not, what about 12 out of 20? The latter would represent a 20% increase in risk over the control rate (12/20 v. 10/20). A true relative risk reduction of 90% may still be plausible, given the observed results and the numbers of patients involved. In short, given this larger number of patients and the lower chance of a "bad sample," the "range of plausible truth" around the observed relative risk reduction of 50% might be narrower, perhaps from a relative risk increase of 20% (represented as –20%) to a relative risk reduction of 90%.

You can develop similar intuitively derived confidence intervals for the larger trials. We've done this in Table 1B, which also shows the 95% confidence intervals that we cal-

Table 1A: Relative risk and relative risk reduction observed in 5 successively larger hypothetical trials

Control event rate	Treatment event rate	Relative risk, %	Relative risk reduction, %*
2/4	1/4	50	50
10/20	5/20	50	50
20/40	10/40	50	50
50/100	25/100	50	50
500/1000	250/1000	50	50

*Calculated as the absolute difference between the control and treatment event rates (expressed as a fraction or a percentage), divided by the control event rate. In the first row in this table, relative risk reduction = (2/4 –1/4) ÷ 2/4 = 1/2 or 50%. If the control event rate were 3/4 and the treatment event rate 1/4, the relative risk reduction would be (3/4 – 1/4) ÷ 3/4 = 2/3. Using percentages for the same example, if the control event rate were 75% and the treatment event rate were 25%, the relative risk reduction would be (75% – 25%) ÷ 75% = 67%.

Table 1B: Confidence intervals (CIs) around the relative risk reduction in 5 successively larger hypothetical trials

Control event rate	Treatment event rate	Relative risk, %	Relative risk reduction, %	CI around relative risk reduction, %	
				Intuitive CI*	Calculated 95% CI*†
2/4	1/4	50	50	–50 to 90	–174 to 92
10/20	5/20	50	50	–20 to 90	–14 to 79.5
20/40	10/40	50	50	0 to 90	9.5 to 73.4
50/100	25/100	50	50	20 to 80	26.8 to 66.4
500/1000	250/1000	50	50	40 to 60	43.5 to 55.9

*Negative values represent an increase in risk relative to control. See text for further explanation.
†Calculated by statistical software.

culated using a statistical program called StatsDirect (available commercially through www.statsdirect.com). You can see that in some instances we intuitively overestimated or underestimated the intervals relative to those we derived using the statistical formulas.

The bottom line

Confidence intervals inform clinicians about the range within which the true treatment effect might plausibly lie, given the trial data. Greater precision (narrower confidence intervals) results from larger sample sizes and consequent larger number of events. Statisticians (and statistical software) can calculate 95% confidence intervals around any estimate of treatment effect.

Tip 2: Interpreting confidence intervals

You should now have an understanding of the relation between the width of the confidence interval around a measure of outcome in a clinical trial and the number of participants and events in that study. You are ready to consider whether a study is sufficiently large, and the resulting confidence intervals sufficiently narrow, to reach a definitive conclusion about recommending the therapy, after taking into account your patient's values, preferences and circumstances.

The concept of a minimally important treatment effect proves useful in considering the issue of when a study is large enough and has therefore generated confidence intervals that are narrow enough to recommend for or against the therapy. This concept requires the clinician to think about the smallest amount of benefit that would justify therapy.

Consider a set of hypothetical trials. Fig. 1A displays the results of trial 1. The uppermost point of the bell curve is the observed treatment effect (the point estimate), and the tails of the bell curve represent the boundaries of the 95% confidence interval. For the medical condition being investigated, assume that a 1% absolute risk reduction is the smallest benefit that patients would consider to outweigh the downsides of therapy.

Given the information in Fig. 1A, would you recommend this treatment to your patients if the point estimate represented the truth? What if the upper boundary of the confidence interval represented the truth? Or the lower boundary?

For all 3 of these questions, the answer is yes, provided that 1% is in fact the smallest patient-important difference. Thus, the trial is definitive and allows a strong inference about the treatment decision.

In the case of trial 2 (see Fig. 1B), would your patients choose to undergo the treatment if either the point estimate or the upper boundary of the confidence interval represented the true effect? What about the lower boundary? The answer regarding the lower boundary is no, because the effect is *less* than the smallest difference that patients would consider large enough for them to undergo the treatment. Al-

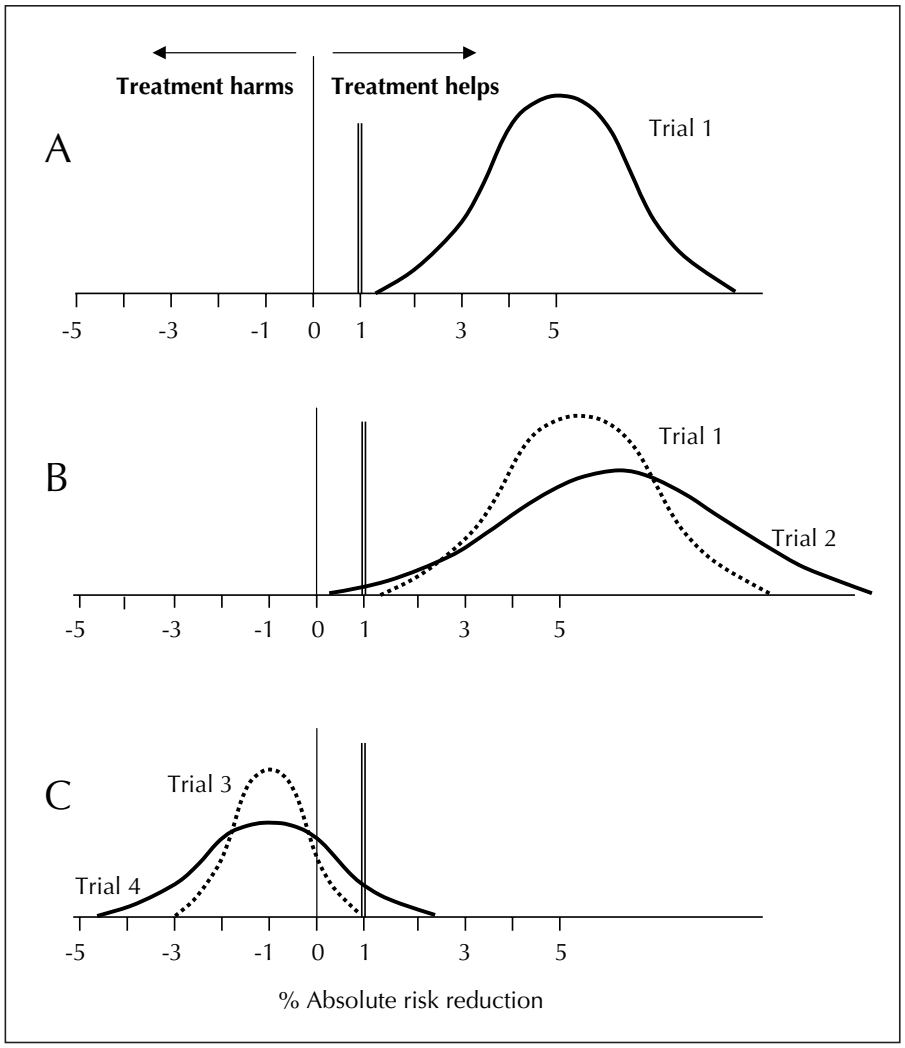

Fig. 1: Results of 4 hypothetical trials. For the medical condition under investigation, an absolute risk reduction of 1% (double vertical rule) is the smallest benefit that patients would consider important enough to warrant undergoing treatment. In each case, the uppermost point of the bell curve is the observed treatment effect (the point estimate), and the tails of the bell curve represent the boundaries of the 95% confidence interval. See text for further explanation.

though trial 2 shows a "positive" result (i.e., the confidence interval does not encompass zero), the sample size was inadequate and the result remains compatible with risk reductions below the minimal patient-important difference.

When a study result is positive, you can determine whether the sample size was adequate by checking the lower boundary of the confidence interval, the smallest plausible treatment effect compatible with the results. If this value is greater than the smallest difference your patients would consider important, the sample size is adequate and the trial result definitive. However, if the lower boundary falls below the smallest patient-important difference, leaving patients uncertain as to whether taking the treatment is in their best interest, the trial is not definitive. The sample size is inadequate, and further trials are required.

What happens when the confidence interval for the effect of a therapy includes zero (where zero means "no effect" and hence a negative result)?

For studies with negative results — those that do not exclude a true treatment effect of zero — you must focus on the other end of the confidence interval, that representing the largest plausible treatment effect consistent with the trial data. You must consider whether the upper boundary of the confidence interval falls below the smallest difference that patients might consider important. If so, the sample size is adequate, and the trial is definitively negative (see trial 3 in Fig. 1C). Conversely, if the upper boundary exceeds the smallest patient-important difference, then the trial is not definitively negative, and more trials with larger sample sizes are needed (see trial 4 in Fig. 1C).

The bottom line

To determine whether a trial with a positive result is sufficiently large, clinicians should focus on the lower boundary of the confidence interval and determine if it is greater than the smallest treatment benefit that patients would consider important enough to warrant taking the treatment. For studies with a negative result, clinicians should examine the upper boundary of the confidence interval to determine if this value is lower than the smallest treatment benefit that patients would consider important enough to warrant taking the treatment. In either case, if the confidence interval overlaps the smallest treatment benefit that is important to patients, then the study is not definitive and a larger study is needed.

Table 2: The 3/*n* rule to estimate the upper limit of the 95% confidence interval (CI) for proportions with 0 in the numerator

n	Observed proportion	3/*n*	Upper limit of 95% CI
20	0/20	3/20	0.15 or 15%
100	0/100	3/100	0.03 or 3%
300	0/300	3/300	0.01 or 1%
1000	0/1000	3/1000	0.003 or 0.3%

Tip 3: Estimating confidence intervals for extreme proportions

When reviewing journal articles, readers often encounter proportions with small numerators or with numerators very close in size to the denominators. Both situations raise the same issue. For example, an article might assert that a treatment is safe because no serious complications occurred in the 20 patients who received it; another might claim near-perfect sensitivity for a test that correctly identified 29 out of 30 cases of a disease. However, in many cases such articles do not present confidence intervals for these proportions.

The first step of this tip is to learn the "rule of 3" for zero numerators,[7] and the next step is to learn an extension (which might be called the "rule of 5, 7, 9 and 10") for numerators of 1, 2, 3 and 4.[8]

Consider the following example. Twenty people undergo surgery, and none suffer serious complications. Does this result allow us to be confident that the true complication rate is very low, say less than 5% (1 out of 20)? What about 10% (2 out of 20)?

You will probably appreciate that if the true complication rate were 5% (1 in 20), it wouldn't be that unusual to observe no complications in a sample of 20, but for increasingly higher true rates, the chances of observing no complications in a sample of 20 gets increasingly smaller.

What we are after is the upper limit of a 95% confidence interval for the proportion 0/20. The following is a simple rule for calculating this upper limit: if an event occurs 0 times in *n* subjects, the upper boundary of the 95% confidence interval for the event rate is about 3/*n* (Table 2).

You can use the same formula when the observed proportion is 100%, by translating 100% into its complement. For example, imagine that the authors of a study on a diagnostic test report 100% sensitivity when the test is performed for 20 patients who have the disease. That means that the test identified all 20 with the disease as positive and identified none as falsely negative. You would like to know how low the sensitivity of the test could be, given that it was 100% for a sample of 20 patients. Using the 3/*n* rule

Table 3: Method for obtaining an approximation of the upper limit of the 95% CI*

Observed numerator	Numerator for calculating approximate upper limit of 95% CI
0	3
1	5
2	7
3	9
4	10

*For any observed numerator listed in the left hand column, divide the corresponding numerator in the right hand column by the number of study subjects to get the approximate upper limit of the 95% CI. For example, if the sample size is 15 and the observed numerator is 3, the upper limit of the 95% confidence interval is approximately 9 ÷ 15 = 0.6 or 60%.

for the proportion of false negatives (0 out of 20), we find that the proportion of false negatives could be as high as 15% (3 out of 20). Subtract this result from 100% to obtain the lower limit of the 95% confidence interval for the sensitivity (in this example, 85%).

What if the numerator is not zero but is still very small? There is a shortcut rule for small numerators other than zero (i.e., 1, 2, 3 or 4) (Table 3).

For example, out of 20 people receiving surgery imagine that 1 person suffers a serious complication, yielding an observed proportion of 1/20 or 5%. Using the corresponding value from Table 3 (i.e., 5) and the sample size, we find that the upper limit of the 95% confidence interval will be about 5/20 or 25%. If 2 of the 20 (10%) had suffered complications, the upper limit would be about 7/20, or 35%.

The bottom line

Although statisticians (and statistical software) can calculate 95% confidence intervals, clinicians can readily estimate the upper boundary of confidence intervals for proportions with very small numerators. These estimates highlight the greater precision attained with larger sample sizes and help to calibrate intuitively derived confidence intervals.

Conclusions

Clinicians need to understand and interpret confidence intervals to properly use research results in making decisions. They can use thresholds, based on differences that patients are likely to consider important, to interpret confidence intervals and to judge whether the results are definitive or whether a larger study (with more patients and events) is necessary. For proportions with extremely small numerators, a simple rule is available for estimating the upper limit of the confidence interval.

This article has been peer reviewed.

From the Department of Medicine, Mayo Clinic College of Medicine, Rochester, Minn. (Montori); the Hospital Medicine Unit, Division of General Medicine, Emory University, Atlanta, Ga. (Kleinbart); the Departments of Epidemiology and Biostatistics and of Pediatrics, University of California, San Francisco, San Francisco, Calif. (Newman); 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); the Department of Pediatrics, University of Texas, Houston, Tex. (Moyer); and the Departments of Medicine and of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ont. (Guyatt)

Competing interests: None declared.

Contributors: Victor Montori, as principal author, decided on the structure and flow of the article, and oversaw and contributed to the writing of the manuscript. Jennifer Kleinbart reviewed the manuscript at all phases of development and contributed to the writing of tip 1. Thomas Newman developed the original idea for tip 3 and reviewed the manuscript at all phases of development. 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. Virginia Moyer reviewed and revised the final draft of the manuscript to improve clarity and style. Gordon Guyatt developed the original ideas for tips 1 and 2, reviewed the manuscript at all phases of development, contributed to the writing as coauthor, and reviewed and revised the final draft of the manuscript to achieve accuracy and consistency of content as general editor.

References

1. Barratt A, Wyer PC, Hatala R, McGinn T, Dans AL, Keitz S, et al. 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.
2. 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.
3. Guyatt G, Walter S, Cook D, Jaeschke R. Therapy and understanding the results: confidence intervals. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual of evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 339-49.
4. 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 [editorial]. *CMAJ* 2004;171(4):347-8.
5. Guyatt G, Montori V, Devereaux PJ, Schunemann H, Bhandari M. Patients at the center: in our practice, and in our use of language. *ACP J Club* 2004;140:A11-2.
6. Jaeschke R, Guyatt G, Barratt A, Walter S, Cook D, McAlister F, et al. Measures of association. In: Guyatt G, Rennie D, editors. *Users' guides to the medical literature: a manual of evidence-based clinical practice*. Chicago: AMA Press; 2002. p. 351-68.
7. Hanley J, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* 1983;249:1743-5.
8. Newman TB. If almost nothing goes wrong, is almost everything all right? [letter]. *JAMA* 1995;274:1013.

Correspondence to: *Dr. Peter C. Wyer, 446 Pelhamdale Ave., Pelham NY 10803, USA; fax 212 305-6792; pwyer@worldnet.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