# Tips for teachers of evidence-based medicine: 2. Confidence intervals and *p* values

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

Clinicians need to understand both *p* values and confidence intervals to be able to use the literature and apply published results to their patients. These statistical concepts are among the most challenging for clinicians learning to use the results of research to guide their clinical care. Such learners are characteristically intimidated by presentations involving complex mathematical formulas and may even have decided that “this is something I can never really understand.”

Hypothesis testing (using *p* values) and estimation (using confidence intervals) are 2 statistical frameworks that pertain to the precision of the observed magnitude of a treatment effect coming from a randomized trial (or a study of weaker design). Hypothesis testing answers the question of whether the treatment is likely to have a non-zero effect with “yes” or “no.” The strategy involves establishing a null hypothesis (typically, that the treatment effect is zero) and determining whether the data suggest that the null hypothesis is sufficiently unlikely that we can reject it. Estimation techniques identify the range within which the effect of a treatment may plausibly lie, given the result we have observed.

This article presents 4 approaches to helping clinicians understand these concepts. As with other articles in this series, clinical educators experienced in teaching evidence-based medicine developed the tips and have used them extensively in teaching the principles of critical appraisal to clinicians and clinical trainees.[1] We have attempted to capture the interactive process that the educators who developed the approaches characteristically observe when using them. We have also emphasized the stumbling blocks that those educators have most commonly encountered among their learners in doing so. 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 4 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. The first 3 tips demonstrate how to introduce the learner to confidence intervals, describe the role of confidence intervals in clinical decision-making and present a shortcut for estimating confidence intervals around results when either very few or almost all subjects experienced the outcome event of interest. The fourth tip offers a formula-free approach to interpreting *p* values when confidence intervals are not provided.

We present the 3 tips related to confidence intervals first because we believe that confidence intervals are more useful for clinicians than *p* values. Indeed, as educators, we avoid teaching about *p* values unless questions from the learners necessitate our doing so.

The opportunity to discuss confidence intervals typically arises as you are reviewing an article with a group of learners. You come to the results, which include a point estimate of the treatment effect and its associated confidence interval. You ask the learners to offer their notions of what the confidence interval means and, after they have made their usually tentative explanations, ask if they would like to spend some time gaining a deeper understanding of the concept of confidence intervals. If they answer in the affirmative, you launch into the first tip. For a particularly enthusiastic group, you may follow tips 1 through 4 in sequence in the same session.

## Teaching tip 1: Making confidence intervals intuitive

### *When to use this tip*

This tip is suitable for clinical learners with a beginner or intermediate level of critical appraisal skills; specifically, they should already have some familiarity with the concepts of relative risk and relative risk reduction.[2] The exercise takes 10 to 15 minutes. The general objective is to help learners overcome the stumbling blocks associated with understanding confidence intervals — fear of statistical concepts (and jargon) and mathematical formulas — with the following specific objectives:

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

### 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/171/6/611)
- An interactive version of this article, as well as other tools and resources, is available at www.ebmtips.net/ci001.asp

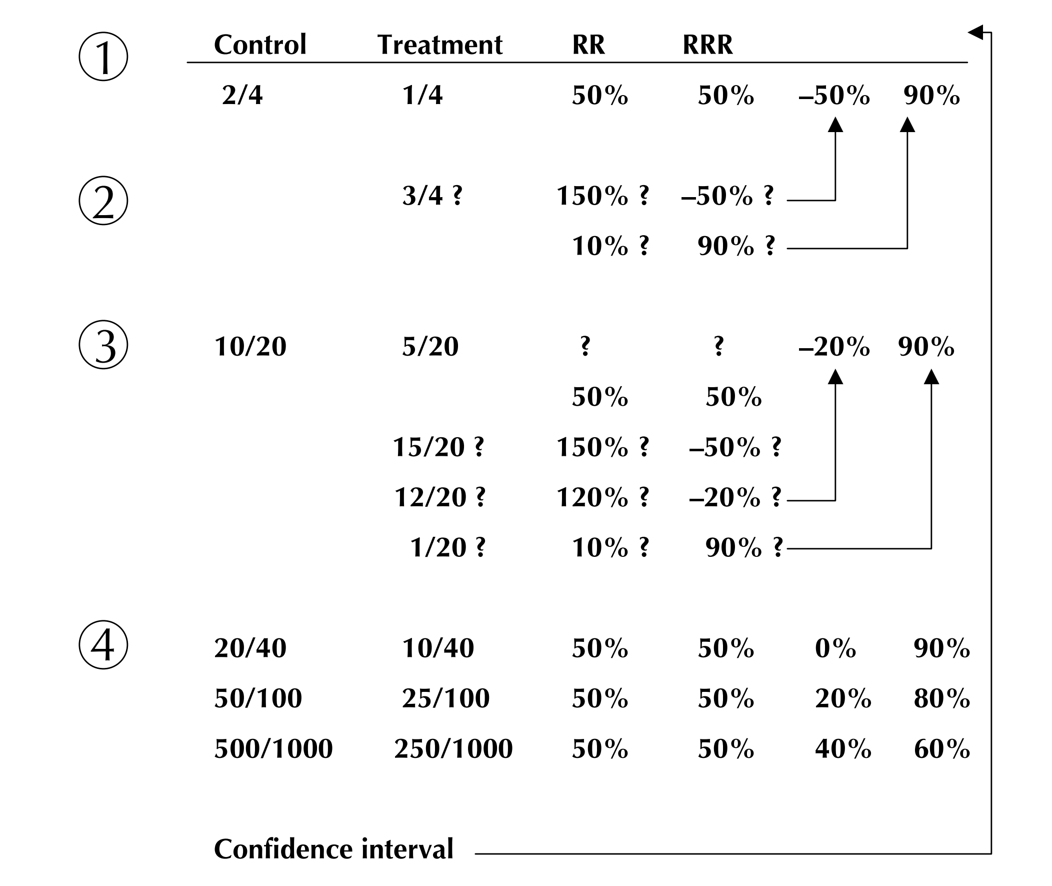
### *The script*

Tell the learners you are going to present the results of a series of experiments. Create a 6-column table with the following column headings: Control (the event rate in the control group), Treatment (the event rate in the treatment group), RR (for relative risk), RRR (for relative risk reduction) and no headings for the final 2 columns (see Fig. 1, step 1). In the first 2 columns, present the event rates as 2/4 and 1/4. Have the learners generate the relative risk (RR) and the relative risk reduction (100 – RR) (both 50%) associated with these results, and write them in their respective columns.

At this point you ask, "Who would be ready to recommend this treatment to a patient?" Usually all students express a strong reluctance to recommend the treatment. On the rare occasion when 1 or 2 learners do not share the group's reluctance, they are quickly brought around to a greater degree of skepticism by discussion with their colleagues. When the group has arrived at a consensus, you challenge them: "Why not give the treatment? Is this not a large relative risk reduction?" At this point, the learners generally make reference to the sample size being too small. You challenge them again: "The sample size is too small? So what?" The response is usually "We do not trust the results." You can then paraphrase, suggesting that what they are really saying is that the true effect may be much larger, or much smaller, than the effect observed.

You can now go further, suggesting that these results may in fact be compatible with the treatment being harmful. You ask, "Is it plausible that, given these results, the true treatment effect is really a 50% *increase* in relative risk? In other words, is it plausible that the true treatment event rate was 3/4 instead of 1/4?" (Fig. 1, step 2). After the learners agree that this large harmful effect may represent the underlying truth, you ask whether a relative risk reduction of 90% would be consistent with the data, and they quickly reply in the affirmative. In the final unlabelled columns, you then write "–50%" and "90%."

You then say that the investigators, on the basis of their preliminary results, have obtained a small grant and can test the treatment on a larger group of patients. In this study, they observe (and you write on the board) event rates of 10/20 for control and 5/20 for treatment (Fig. 1,



| | Control | Treatment | RR | RRR | | |
|---|---|---|---|---|---|---|
| ① | 2/4 | 1/4 | 50% | 50% | –50% | 90% |
| ② | | 3/4 ? | 150% ? | –50% ? | | |
| | | | 10% ? | 90% ? | | |
| ③ | 10/20 | 5/20 | ? | ? | –20% | 90% |
| | | | 50% | 50% | | |
| | | 15/20 ? | 150% ? | –50% ? | | |
| | | 12/20 ? | 120% ? | –20% ? | | |
| | | 1/20 ? | 10% ? | 90% ? | | |
| ④ | 20/40 | 10/40 | 50% | 50% | 0% | 90% |
| | 50/100 | 25/100 | 50% | 50% | 20% | 80% |
| | 500/1000 | 250/1000 | 50% | 50% | 40% | 60% |

**Fig. 1: Graphic illustration of tip 1, for the intuitive calculation of the confidence interval.** Each circled number to the left of the illustration represents a successive step in this exercise. Data for 5 successively larger hypothetical trials with the same observed relative risk and relative risk reduction, used for intuitive generation of confidence intervals. See text for further explanation.

step 3). You ask, "Is it still plausible that the underlying truth could be a relative risk increase of 50%? That is, might it be that the treatment event rate is really 15/20 instead of 5/20?" A group member may say, "It's possible," in which case you ask, "But does it remain plausible?" When the group rejects this estimate of the underlying truth, you question whether an increase in relative risk of 20% remains a reasonable possibility (that is, would it be plausible to observe an event rate of 5/20 when the truth is actually 12/20?). Most will agree, and you write "–20%" below the "–50%" from the previous scenario. You then ask, "Is it plausible that the relative risk reduction is 90%? That is, is it plausible that the treatment event rate is 1/20 instead of the observed 5/20?" They agree, and you write "90%" in the last column.

You tell the learners that the investigators now have some very tantalizing results and are able to obtain a larger grant to support yet another experiment. You repeat the exercise successively, maintaining the same relative risk and relative risk reduction, but with increasing sample sizes, generating results of 20/40 and 10/40, 50/100 and 25/100, 500/1000 and 250/1000 (Fig. 1, step 4). The students typically generate boundaries of plausible truth for the relative risk reductions in the vicinity of 0 and 90%, 20% and 80%, and 40% and 60% for the progressively larger experiments.

At the conclusion of the exercise you inform the group that they have, intuitively, generated a confidence interval around each result, and you add that label to cover the last 2 columns in Fig. 1. You point out that, with identical point estimates of relative risk and relative risk reduction, the boundaries around the plausible truth have narrowed as the sample size has increased.

At the same time, you note the inevitable discomfort and disagreement that group members have experienced as they generated these boundary values. Fortunately, you inform them, we don't have to rely on our intuition, but can ask a statistician (or statistical software) to calculate a 95% confidence interval. If we wish to be more conservative, we can ask for a 99% confidence interval. You conclude by reiterating that, absent strong data from other studies to the contrary, the point estimate provides the single most likely estimate of the underlying true treatment effect, and the confidence interval represents the range within which that true effect plausibly lies.

Some care is needed in describing the 95% confidence interval to learners unfamiliar with the concept. In particular, it is not strictly accurate to say that the 95% confidence interval represents the range in which it is 95% likely that the true effect lies. A more accurate statement is that, if the relative risk observed in a particular study were the true relative risk, and we repeated the study many times, in 95% of those repetitions, the result would fall within the confidence interval. Teachers can decide for themselves whether it is important to use the precisely correct statement, or whether there is an appreciable advantage in explaining the confidence interval in terms of the boundaries of probable truth.

### *The bottom line*

- Stumbling blocks: Fear of statistical concepts (and jargon) and mathematical formulas.
- The instructor presents the event rates for the control and treatment groups, and the group computes the relative risk and the relative risk reduction and then generates an intuition-based confidence interval.
- This process involves minor calculations and an understanding of risk measures.
- The process of defining confidence intervals is inductive — the concept is derived and demonstrated before it is given a label. This overcomes the common stumbling block of learners' fears of technical terms which may distract them from the underlying concepts.
- No background in statistics is required for the learners to discover the dynamic relation between sample size and confidence intervals.

See Appendix 1 for the summary card for this tip.

## Teaching tip 2: Interpreting confidence intervals

### *When to use this tip*

This tip is suitable for highly committed clinicians and clinical trainees who have an intermediate to advanced level of understanding of the principles of critical appraisal. An understanding of the concept of minimally important treatment effect is a prerequisite for this approach. The exercise usually takes 15 to 20 minutes. The general objective of this tip is to provide clinicians with a way of deciding whether a study is sufficiently large and the resulting confidence intervals sufficiently narrow, with the following specific objective:

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

Learners who have grasped the basic principle of the confidence interval can quickly decide if it excludes a true treatment effect of zero by answering the following questions: Does the confidence interval for relative risk include 1? Does the confidence interval for relative risk reduction and absolute risk reduction include zero? Does the confidence interval for number-needed-to-treat include infinity? However, learners may stumble when trying to come to additional conclusions about the implications for patient management. In particular, they may have difficulty when challenged to draw conclusions about whether the study results support a strong inference that the treatment should or should not be administered.

This tip provides an answer to the question, "How do you decide if the sample size is large enough?" Often, when the instructor poses this question to the learners, a more sophisticated group member will raise issues of power analysis and α and β error. Although an understanding of power analysis is important for investigators planning a clinical trial, it is gratuitous for clinicians interpreting a trial's results. In the course of this exercise, learners discover that clinicians can determine if a study's sample size is sufficiently large by evaluating the relation between the boundaries of the confidence interval and the minimal treatment effect that is important to patients (understood as the treatment effect above which a treatment's benefits outweigh its harms and costs and below which they do not).[3]

## *The script*

Draw and label the axes of Fig. 2, step 1. As you draw the elements of the figure, label them out loud: "The vertical line in the centre represents the line of no difference in event rate

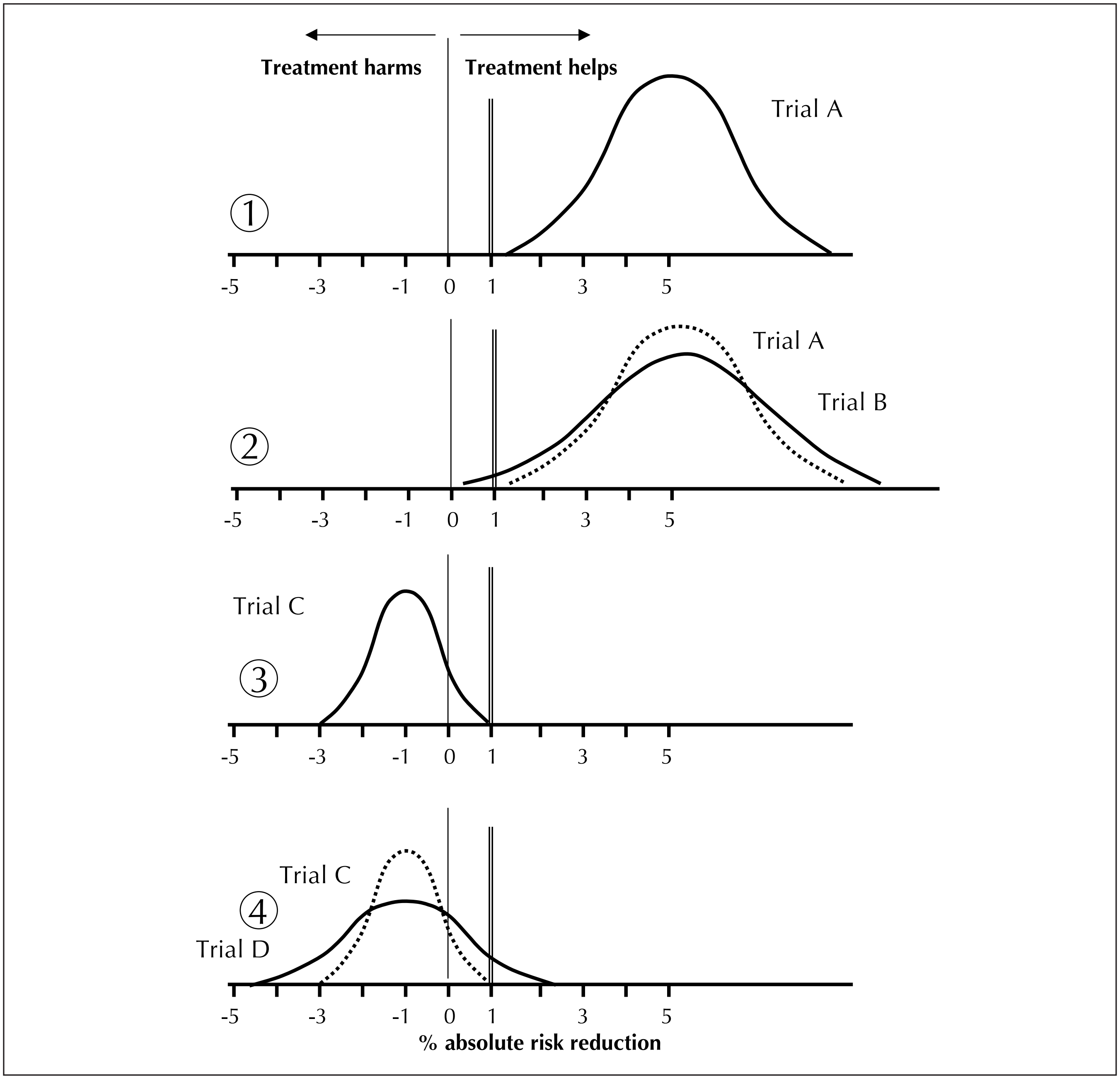


**Fig. 2: Graphic illustration of tip 2, for the assessment of clinical significance and sample size using confidence intervals.** Each circled number to the left of the illustration represents a successive step in the exercise. The figure shows the results of 4 hypothetical trials. For the medical condition 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.

between the treatment and control groups, an absolute risk reduction of zero. To the right of the line we will find studies in which the treated group had a lower event rate than the control group. To the left of the line we will find studies in which the treated group had a higher event rate than the control group. Given the risks, costs and benefits of these 2 treatments, let us assume that absolute risk reductions in the event rate of 1% (Fig. 2, step 1, double vertical line) or more warrant treatment, whereas reductions of less than 1% do not warrant treatment, because the risks and costs outweigh the benefits."

Now, draw trial A (Fig. 2, step 1), in which the observed treatment effect (sometimes referred to as the point estimate of the treatment effect) is an absolute risk reduction in the event rate of 5% (that is, 5% more of the patients in the control group died than in the treatment group). Notice that you will be using a bell-shaped curve to depict the distribution of results likely to be observed in a large number of trials of similar design and size. The point estimate (at the centre of the distribution) represents the most likely true value. As values depart from the point estimate they become less likely to represent the true value. The 95% confidence interval is represented by the width of the curve. Values outside of this, represented by the tails of the curve, are significantly less likely to represent the true value. Ask the learners, "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? What about the lower boundary?" For all 3 markers, the answer will likely be yes, because they all fall above the threshold of the smallest patient-important difference. Thus, the trial is definitive and allows a strong inference about the treatment decision.

Next, draw the data for trial B, which, like trial A, should be centred at 5% absolute risk reduction, but with a wider confidence interval (Fig. 2, step 2). These data depict the results of another trial with a smaller sample size than that of trial A. Ask, "Would you recommend this treatment for your patients if the true effect were represented by the point estimate or the upper boundary of the confidence interval?" The learners will answer in the affirmative. "What about the lower boundary?" The answer here will be no, for the effect is less than the smallest difference that we have decided warrants treatment. At this point you lead the learners to the conclusion that although trial B shows a positive result (the confidence interval excludes an effect of zero), the sample size was inadequate and yielded a result that is still compatible with risk reductions below the minimal patient-important treatment effect. Thus, while the learners' patients may still consider taking the treatment, the inference is not nearly as strong.

Suggest to your learners that for studies with positive results, they should determine if the sample size was adequate by checking the lower boundary of the confidence interval. If this lower boundary, the smallest plausible treatment effect compatible with the results, is greater than the smallest patient-important difference, the sample size is adequate and the positive trial result definitive. If the lower boundary falls below the smallest patient-important difference, the trial is not definitive, the sample size is too small, and further trials are required.

Now, draw the data for trial C, explaining that "this is the result of a trial that yielded a point estimate of 1% *increase* in the risk of the event of interest with treatment relative to control" (Fig. 2, step 3). Once again, you ask students if they would recommend the treatment if the point estimate or the upper or lower boundary of the confidence interval represented the true effect. In each case, the answer will be no, for the effect of treatment would either be harmful or less than the 1% absolute benefit we have stipulated as the smallest patient-important difference. This trial has therefore excluded any patient-important benefit and can be considered definitive.

Repeat the exercise with the data for trial D as shown in Fig. 2, step 4. Trial D, which also had a negative result, represents a situation analogous to trial B, in which the confidence interval is wide because of small sample size. Here, if the upper boundary of the confidence interval actually represented the truth, we would recommend the treatment. Thus, the trial does not exclude a patient-important benefit of treatment and cannot be considered definitive. Suggest that for studies with negative results, learners should consider whether the upper boundary of the confidence interval, the largest treatment effect compatible with the data, is less than the smallest difference that patients would consider important. If so, the sample size is adequate, and the trial is definitively negative. If the upper boundary exceeds the smallest patient-important difference, then the trial is not definitive, and more trials with larger sample sizes are needed.

## Extension for advanced learners

You can use this example to illustrate an additional concept. Sometimes, investigators test an experimental treatment not because they expect it to yield greater benefit, but rather because it is less costly, less toxic or less inconvenient. Laparoscopic or day surgery, less aggressive chemotherapy regimens and less intense anticoagulation for antithrombotic prophylaxis are examples of interventions that might be tested in so-called "equivalence" trials.

Ask the learners to assume that trial C in Fig. 2 represents the results of such a study and that negative values of absolute risk reduction (i.e., to the left of the line representing zero absolute risk reduction) favour a less toxic experimental chemotherapy regimen. The smallest absolute risk reduction in favour of the more toxic drug that would make it preferable to patients is (positive) 1%, i.e., to the right of the line representing zero absolute risk reduction. In trial C, the less toxic drug has apparently reduced the mortality rate by 1% compared with the more toxic drug. Are the learners ready to confidently substitute the new regimen? Their answer is yes. Although the results do not show that the experimental treatment is definitively superior in reducing mortality, the confidence interval does exclude the smallest mortality difference in favour of the conventional, more toxic regimen that patients would consider important. What about the results in trial D? Here, one could argue in favour of continuing the conventional regimen, for the results have not excluded mortality benefits greater than 1% in favour of the more toxic treatment.

### *The bottom line*

- Stumbling blocks: Inability to determine if a study provides definitive evidence of efficacy or of lack of efficacy, that is, difficulty in deciding if the sample size was large enough.
- This tip allows learners to discover the importance of analyzing the boundaries of the confidence intervals in studies with positive and negative results to determine if the results are definitive.

See Appendix 1 for the summary card for this tip.

## Teaching tip 3: Estimating confidence intervals for extreme proportions

### *When to use this tip*

This tip is suitable for intermediate-level learners, and the exercise takes 5 to 10 minutes. The general objective is to introduce learners to a shortcut method of estimating confidence intervals for extreme proportions, with the following specific objectives:

- Learn to estimate confidence intervals for proportions with very low numerators.
- Learn to estimate confidence intervals for proportions with numerators very close to the denominators.

This tip is suitable for learners who know what confidence intervals are, or it can follow the scripts presented previously in this article. It addresses the fairly common situation that readers encounter in reviewing journal articles: proportions with small numerators or with numerators very close in size to the denominators. These 2 situations raise the same issues. 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. Motivated learners may be interested in overcoming this limitation of the literature but may be unable or unwilling to use paper and pencil or software to calculate the upper boundary of the confidence interval for a proportion with a low numerator.

The first step of this tip is to learn the "rule of 3" for zero numerators,[4] 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.[5]

### *The script*

When discussing an article reporting an observed proportion of 0%, ask the group how sure they are that the proportion being measured is actually close to 0%. For example, you might say, "Twenty people have undergone surgery, and none suffered serious complications. [At this point, write "0/20" on the board.] Does that allow us to be confident that the true complication rate is very low, say less than 5% (1 out of 20)? Are we pretty sure it's less than 10% (2 out of 20)?" Most learners will appreciate that if the true complication rate is 5% (1 in 20), it wouldn't be that unusual to observe no complications in a sample of 20, but as the true rate increases, the chances of observing no complications in a sample of 20 gets progressively smaller.

This leads to a discussion of confidence intervals. You can say, "So we'd like to know how high the true rate might plausibly be, given our observed rate of 0 out of 20." That number is the upper limit of a 95% confidence interval for the proportion 0/20. After the learners provide some intuitive estimates, offer the following shortcut: if an event occurs 0 times in *n* trials (or occurs 0 times among *n* patients), the upper boundary of the 95% confidence interval for the event rate is about 3/*n*.

Draw a 2-column table on the board and label the columns "Observed proportion (0/*n*)" and "Upper limit of 95% CI (3/*n*)," where CI stands for "confidence interval." In the first row of the left column, write "0/100" and have the group generate the entry for the next column: "3/100." Then enter "0/300" in the second row of the first column and have the learners generate "3/300." The last step is to nail down this shortcut by having them convert the entries in the second column to the corresponding percentage. "So if you observe 0 out of 100, the true rate is probably less than what percent? What if you observe 0 out of 300? Zero out of 1000? Zero out of 20?" Enter the answers in a third column, labelling it, "Upper limit of the 95% CI." This column is, of course, identical with the second column, with the fractions simplified to decimal form and then converted to a percentage.

You can have the same discussion when the observed proportion is 100%, by translating 100% into its complement. For example, you might say, "Investigators studying the performance of a diagnostic test for a particular disease did the test on 20 patients known to have the disease and observed 100% sensitivity. That means they had no false negatives. How high could the false negative rate be? If you knew that, you could subtract it from 100 and obtain the lower limit of the confidence interval around the sensitivity." When they have grasped the concept, learners will generate boundaries of plausible false negatives of 15% for 0/20, 3% for 0/100, and 0.3% for 0/1000.

After the participants have become familiar with the "rule of 3," you can move on to the shortcut for estimating the upper limit of the confidence interval for proportions with low numerators. You can say something like, "In this study, the investigators observed only 1 complication among the 20 patients. Does anyone know what the upper limit of the 95% confidence interval for the complication rate would be?" If no one knows the answer, you can back up and make sure people can apply the rule of 3 by asking, "What if there had been zero complications?" Once you establish that if no complications had been observed, the upper limit of the 95% confidence interval would be about 3/20, or 15%, you should be able to get the group to say that the upper limit of the confidence interval for 1/20 should be something higher than 15%.

You can then point out that there is a shortcut for low numerators other than zero (i.e., 1, 2, 3 and 4). Then simply present the 2 columns shown in Table 1 on the board. Now illustrate the use of this shortcut interactively, writing various examples on the board. For example, for the observed proportion of 1/20, the upper limit of the 95% confidence interval is approximately 5/20 or 25% (probably close to what people might have guessed). For an observed proportion of 2/20, the upper limit of the 95% confidence interval would be approximately 7/20 or 35%. This method gives a rough estimate of the upper limit of the 95% confidence interval that is more accurate if the denominator is 20 or higher. For example, using the shortcut, the upper limit of the 95% confidence interval for 4/20 is about 10/20 or 50%, whereas the exact value, using the standard statistical formula, is 44%. The upper limit of the 95% confidence interval for 4/100 is about 10/100 or 10%, whereas the exact value is 13%.

### *The bottom line*

- Stumbling block: Uncertainty about the level of confidence in an event rate that is close to 0% or close to 100% when the investigators have not provided a confidence interval.
- This tip can be used in conjunction with tip 2 to restate the importance of analyzing the boundaries of the confidence interval.
- This tip gives learners a practical shortcut to use when reading journal articles.
- In conjunction with the previous tips, this tip illustrates how increasing the sample size narrows the confidence interval.

**Table 1: Method for obtaining an approximation of the upper limit of the 95% confidence interval (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, the learner substitutes the numerator in the right hand column. When this value is divided by the number of study subjects, the learner obtains a reasonable approximation of the 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%.

See Appendix 1 for the summary card for this tip.

## Teaching tip 4: The *p* value

### *When to use this tip*

This tip is suitable for clinical learners who have a beginner or intermediate level of comfort with critical appraisal, and the exercise should take 10 to 15 minutes. The general objective is to introduce learners to the interpretation of the *p* value, with the following specific objectives:

- Understand and interpret the *p* value.
- Understand the limitations of the *p* value.

When you and your learners are working on understanding the results of a paper, you come across *p* values. Your learners seem interested in discussing this statistical term. The stumbling block is that learners' intuitive interpretations of *p* values may have little relation to what *p* values actually signify. Some learners describe a *p* value as the probability of the hypothesis of no effect being correct, given the data. Others believe that a *p* value represents the probability of making a mistake if one rejects the hypothesis of no effect, given the data. Others may invoke type I and type II error, α and β, and other statistical terms without clarity or understanding. This tip avoids statistical labels and instead uses inductive reasoning to derive concepts related to generating theoretical distributions of results, assuming that a null hypothesis is true.

Keep in mind that if the prompt to review *p* values arose because the paper being analyzed did not publish the confidence intervals around the treatment estimates, one option is to calculate these confidence intervals.[6] An online calculator that can be used to generate confidence intervals for dichotomous outcome data is available (www.healthcare.ubc.ca/calc/clinsig.html). Another option is to point out to the learners that confidence intervals can be roughly estimated by recognizing that a *p* value of 0.05 in connection with a positive observed effect means that the lower limit of the 95% confidence interval coincides with the point of no difference.[7] The following script is intended for use with learners who desire to understand *p* values as they relate to hypothesis testing.

### *The script*

Refer to Fig. 3 throughout this teaching tip. In our practice, steps 1 and 2 of the figure are developed simultaneously. Start by beginning to draw the middle portion of the bar diagram illustrated in Fig. 3 on the board. As you do this, explain that a fair coin is flipped 10 times and that this procedure is then repeated 10 000 times. Tell them that you are plotting the percentage of repetitions that yield a particular number of heads: the taller the bar, the higher the percentage. First, ask the learners in what proportion of the coin flips they expect to see 5 heads and 5 tails. The disparity of the suggestions may surprise you (and should provide a message about why intuitive inference is so dangerous). As you expand the figure, ask the learners to speculate about the proportion of times they would see 6 heads and 4 tails, or 4 heads and 6 tails. You can also ask them how frequently they would expect to see 10 heads or 10 tails and, depending on the group dynamic, one or two of the other possible outcomes.

The next step is to inform them of what will actually happen: in a very large number of coin flips, using a standard statistical formula, they will see 5 heads and 5 tails 25% of the time. Write this on the board as illustrated in Fig. 3. You may immediately add that they would see

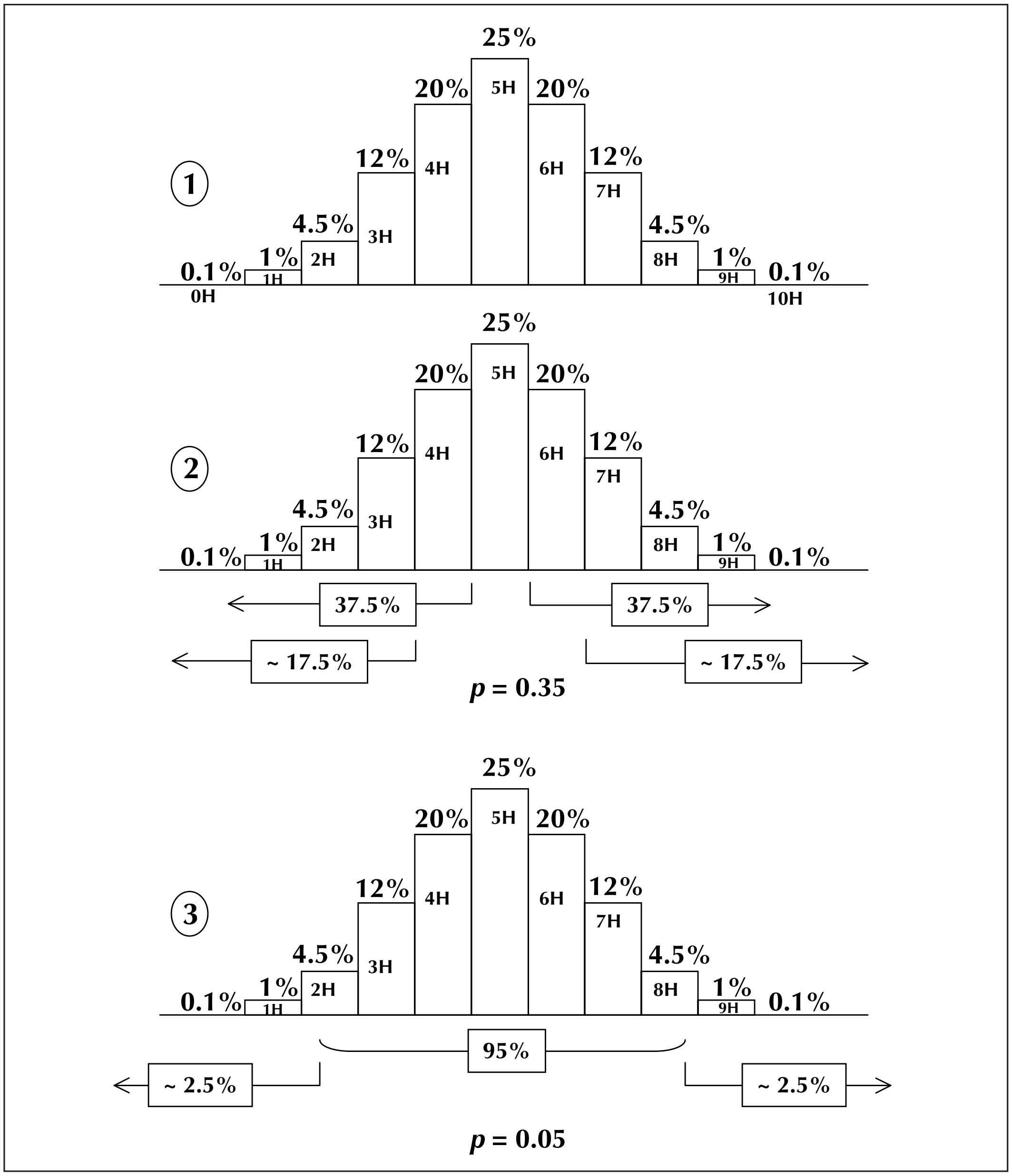


**Fig. 3: Graphic illustration of tip 4, for the calculation of the *p* value.** The circled numbers at the left of the figure refer to the steps described in the text. The chart plots the results of a coin-flipping exercise, in which a fair coin is flipped 10 times and the number of heads is recorded. The procedure is repeated 10 000 times, and each bar in the chart shows the percentage of the repetitions yielding a particular number of heads (1H, 2H, and so on). See text for further explanation.

other proportions, with greater or fewer than 5 heads, 75% of the time. Draw a line from each side of the "5H" bar as you say that half of 75%, or 37.5%, is the probability of proportions with fewer than 5 heads, and the other half, 37.5%, is the probability of proportions with greater than 5 heads (Fig. 3, step 2). Show them that 4 or 6 heads will happen 20% of the time each — a total of 40%. Again, draw a line that encompasses the extremes of the distribution, including 7 heads and 3 heads. Ask them to label those lines as the percentage of times that a result with fewer than 4 heads or greater than 6 heads will be observed (the students may initially fail to take into account the 25% of the 5H bar) and proceed to label those as 17.5% each. Now ask the following question: "If, on a single coin flip series, you observe 7 heads, what is the probability of observing results as extreme or more extreme than the observed 7 heads, simply by chance?" The learners will reply with the correct answer: 35% (they may start with 17.5%, in which case you will remind them of the other side of the distribution). You inform them that this probability, 35%, is the *p* value for observing results more extreme than 7 heads or 3 tails when a fair coin is flipped 10 times.

Explain to the learners that in the context of a clinical trial, the assumption of no treatment effect is analogous to the "fair coin" situation represented in the exercise. Just as they have generated a distribution of the results of a large number of coin flips (assuming an unbiased coin), the statistician generates a theoretical distribution of the possible trial results assuming no difference between treatment and control. The *p* value represents the proportion of that distribution that is as extreme as or more extreme than the observed results (point back at the proportion of the distribution as extreme as or more extreme than the 7/3 split). Return to the figure and show them that the *p* value associated with 10 flips of which 8 are heads is 0.11 and explain how the *p* value becomes smaller as the observed results become more extreme. Suggest that "at one point we will have to agree that the probability of no difference is so small, given the observed results, that the assumption of no difference (the assumption of a 'fair coin') is no longer sustainable, and we reject it. That point is conventionally, and somewhat arbitrarily, set at less than 1 in 40, or less than 2.5%, in either direction." This final point is illustrated in Fig. 3, step 3. Because of the small numbers of coin flips per trial, the point corresponding to a "tail" of 2.5% probability does not coincide exactly with a number in one of the boxes of the histogram. However, at this point in the exercise, learners are usually able to grasp the concept and its relevance to the figure.

### *The bottom line*

- Stumbling blocks: Learners' intuitive interpretations of *p* values may have little relation to what *p* values actually signify, and attempts to understand them have been limited by statistical jargon.
- The learners derive the definition of the *p* value, learning the label only after understanding the concept.
- The tip finishes by giving the learner a feel for the conventional (and arbitrary) nature of the 0.05 threshold *p* value for statistical significance.

See Appendix 1 for the summary card for this tip.

## Report on field-testing

One of the authors (S.K.), an experienced teacher of evidence-based medicine who was not involved in developing the scripts, field-tested the scripts with 15 US medical residents during two 1-hour teaching sessions. Eight of the participants were naive learners with very little experience in evidence-based medicine, and the other 7 were moderately aware of but did not have a strong background in evidence-based medicine.

S.K. took a total of 2 hours to prepare her presentation of the 4 scripts. In that time, she produced transparencies and handouts with the figures. In both sessions she progressed through the 4 tips in the order presented in this article.

Learners not comfortable with the concepts of risk required more detailed guidance to complete the table in tip 1. The more advanced learners wanted to complete the table by calculating absolute risk reduction and number needed to treat, rather than by calculating the relative risk reduction. Some learners demanded the formula for calculating confidence intervals. Very naive learners had difficulty understanding what was meant by the terms "plausible" and "possible" and offered improbable estimates in response to the event rates offered.

Tip 2 was very successful when presented after tip 1, in particular for those who had struggled with the numeric approach used in tip 1. The learners readily grasped how to determine when a trial result was definitive. Tip 3 worked best when the learners worked through a published paper as an example. It was clear during the field test that the learners had to be comfortable with the concepts of risk reduction and number needed to treat.[2] For tip 4, the learners needed to flip some coins to get an idea about the proportions shown in the graph.

When asked about the importance of the concepts and the clarity of the presentations, the learners scored the 4 tips similarly. Tips 1 and 4 received the lowest scores, and tips 2 and 3 the highest scores. The mean scores for all tips ranged from 6.6 to 8.4 on a visual analogue scale from 0 to 10. For learners, the most important messages included the rationale for preferring confidence intervals over *p* values and the interpretation of confidence intervals in the clinical context.

## Conclusions

In this article we have provided scripts for a series of tips that teachers can use to help learners overcome the stumbling blocks associated with the statistical concepts of confidence intervals and *p* values. These tips allow an intuitive approach, so that learners can derive fundamental concepts before applying labels. We have suggested some practical uses for these concepts that should make their understanding desirable to learners.

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); 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, 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. Gordon Guyatt, as general editor, developed the original ideas for tips 1, 2 and 4, 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.

## 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 [editorial]. *CMAJ* 2004;171(4):347-8.
2. Barratt A, Wyer PC, Hatala R, McGinn T, Dans AL, Keitz S, et al. 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/data/171/4/353/DC1/1 (accessed 2004 Sep 7).
3. 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.
4. Hanley J, Lippman-Hand A. If nothing goes wrong, is everything all right? Interpreting zero numerators. *JAMA* 1983;249:1743-5.
5. Newman TB. If almost nothing goes wrong, is almost everything all right? Interpreting small numerators [letter]. *JAMA* 1995;274:1013.
6. Altman D. *Practical statistics for medical research*. London: Chapman and Hall; 1991.
7. 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.

***Correspondence to:*** *Dr. Peter C. Wyer, 446 Pelhamdale Ave., Pelham NY 10803, USA; fax 212 3056792; 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.*

## Appendix 1: Summary cards for 4 teaching tips on confidence intervals and *p* values.

This appendix has been designed so that it can be printed on two sheets of 8 1/2 × 11 inch paper. The individual summary cards can then be cut out, if desired, for use during teaching sessions.

### Teaching tip 1: Making confidence intervals intuitive

**Scenario:** Consider a series of randomized controlled trials that are identical save for an increasing number of subjects. Have the learners decide what magnitudes of relative risk and relative risk reduction are plausible, given the number of subjects in each trial.

1. Create a table with 6 columns. The first 4 columns have the headings “Control” (for the event rate in each control group), “Treatment” (for the event rate in each treatment group), “RR” (for relative risk), “RRR” (for relative risk reduction). Leave the last 2 columns without any heading for the moment.
2. Write “2/4” and “1/4” in the first 2 columns of the first row. The learners complete the entries for relative risk and relative risk reduction and discuss whether they would recommend the new therapy. They also discuss the range of true underlying values of relative risk reduction that could be plausible, given the data, and the upper and lower limits of the range in the last 2 columns (treatment event rates of 3/4 or 0/4 yield relative risk reductions of –50% to +90%).
3. Repeat the exercise with a control event rate of 10/20 and a treatment event rate of 5/20. The learners arrive at a narrower range of plausible relative risk reduction.
4. Repeat with 40, 100 and 1000 patients per arm to generate progressively narrower ranges as the sample size increases.
5. Add a heading for the last 2 columns, “Confidence interval,” and explain that for each row these values represent the range in which the true effect plausibly lies (with 95% confidence, for instance).

**Summary points**

- Stumbling blocks: Fear of statistical concepts (and jargon) and mathematical formulas.
- The instructor presents the event rates for the control and treatment groups, and the group computes the relative risk and the relative risk reduction and then generates an intuition-based confidence interval.
- This process involves minor calculations and an understanding of risk measures.
- The process of defining a confidence interval is inductive: the concept is derived and demonstrated before it is given a label. This overcomes learners’ fears of technical terms.
- No background in statistics is required for learners to discover the dynamic relation between sample size and confidence intervals.

### Teaching tip 2: Interpreting confidence intervals

**Scenario:** The results of a series of 4 trials are represented graphically as distributions of likely results around an observed result to demonstrate the various clinical implications of the confidence limits.

1. Plot a horizontal line representing the range of possible treatment effects (relative risk or risk difference) and a vertical line representing no treatment effect (relative risk of 1, risk difference of 0). Show positive treatment effects to the right of the line and negative effects to the left. Identify the point where risk difference = 1% as the minimal patient-important difference.
2. Draw a normal curve representing the results of many identical studies distributed around a likely true value, with the intersections of the extremes and the horizontal line constituting the confidence interval. The learners decide whether the lower boundary, the point estimate and the upper boundary are greater than the minimal patient-important difference
3. Repeat the exercise as follows:
   a) Point estimate > risk difference of 0, lower boundary of confidence interval > minimal patient-important difference (positive study, strong inference)
   b) Point estimate > risk difference of 0, lower boundary of confidence interval < minimal patient-important difference (positive study, weaker inference).
   c) Point estimate < risk difference of 0, upper boundary of confidence interval < minimal patient-important difference (negative study, strong inference).
   d) Point estimate < risk difference of 0, upper boundary of confidence interval > minimal patient-important difference (negative study, weaker inference).

**Summary points**

- Stumbling blocks: Inability to determine if a study provides definitive evidence of efficacy and inability to use confidence intervals to judge whether the sample size was large enough.
- Learners discover the importance of analyzing the boundaries of the confidence intervals to determine if the results are definitive.

### Teaching tip 3: Estimating confidence intervals for extreme proportions

**Scenario:** A series of simple hypothetical situations that define the problem of estimating the precision of a result when no events or only a few events have been observed. A set of simple numeric exercises is used to solidify comprehension of the rule for determining the 95% confidence interval in such situations.

1. Propose an observed event rate of 0/20. The learners consider successively higher true rates and notice that higher true event rates become less plausible. They are led to the question, "How high can the true event rate be, given an observed event rate of 0/20?"
2. Present the "rule of 3": for event rates of 0/*n*, the upper boundary of the confidence interval is about 3/*n*. Show that this rule can be used for event rates of 100% by subtracting 3/*n* x 100 from 100%.
3. Offer other rules for small numerators:

| Observed numerator | Numerator for calculating upper limit of 95% confidence interval |
|---|---|
| 0 | ~3 |
| 1 | ~5 |
| 2 | ~7 |
| 3 | ~9 |
| 4 | ~10 |

**Summary points**

- Stumbling blocks: Uncertainty about the level of confidence in a result in which the event rate is close to 0% or close to 100% when the investigators have not provided a confidence interval.
- Learners understand the importance of analyzing the boundaries of the confidence interval when no events or few events occur.
- This tip gives learners a practical shortcut to use when reading journal articles.
- Learners understand how increasing the sample size narrows the confidence interval.

### Teaching tip 4: The *p* value

**Scenario:** A hypothetical coin flipping exercise, to convey the concept of the likelihood of event rates greater or less than those expected from a fair coin (no difference in likelihood).

1. Create a histogram with the proportions of 10 000 repetitions of 10 flips of a fair coin that yield the possible distributions from 10 heads, through 5 heads and 5 tails, through 10 tails. Learners are steered to approximations of the correct answers, beginning with the 25% likelihood of 5 heads and 5 tails and continuing through 20% likelihood of 6 heads or of 6 tails, 12% likelihood of 7 heads or of 7 tails, 4.5% likelihood of 8 heads or of 8 tails, 1% likelihood of 9 heads or of 9 tails, and 0.1% likelihood of 10 heads or of 10 tails.
2. The learners determine the probability of 7 or more heads or tails. They determine that they must add the proportions on both sides of the distribution. They are led to recognize that, at some point, the probability of obtaining a particular result, assuming a fair coin, is too low, and that they would have to reject such an assumption. Present the parallel with the *p* value and rejection of the null hypothesis of no treatment effect.

**Summary points**

- Stumbling blocks: Learners' intuitive interpretations of *p* values may have little relation to what *p* values actually signify, and attempts to understand them have been limited by statistical jargon.
- Learners derive the definition of the *p* value, learning the label only after understanding the concept.