

THERAPY AND UNDERSTANDING THE RESULTS

Confidence Intervals

Gordon Guyatt, Stephen Walter, Deborah Cook,
and Roman Jaeschke

The following EBM Working Group members also made substantive contributions to this section: Mark Wilson and Martin Stockler

IN THIS SECTION

How Should We Treat Patients With Heart Failure? A Problem in Interpreting Study Results

Solving the Problem: What Are Confidence Intervals?

Using Confidence Intervals to Interpret the Results of Clinical Trials

Interpreting Apparently “Negative” Trials

Interpreting Apparently “Positive” Trials

Was the Trial Large Enough?

Conclusion



Hypothesis testing involves estimating the probability that observed results would have occurred by chance if a *null hypothesis*, which most commonly states that there is no difference between a treatment condition and a control condition, were true (see Part 2B2, “Therapy and Understanding the Results, Hypothesis Testing”). Health researchers and medical educators have increasingly recognized the limitations of hypothesis testing; consequently, an alternative approach, estimation, is becoming more popular. A number of authors¹⁻⁵ have outlined the concepts that we will introduce here, and you can use the full expanse of their discussions to supplement our presentation. We will illustrate the concepts with an example introduced earlier in this book (see Part 2B2, “Therapy and Understanding the Results, Hypothesis Testing”).

HOW SHOULD WE TREAT PATIENTS WITH HEART FAILURE? A PROBLEM IN INTERPRETING STUDY RESULTS

In a double-blind randomized controlled trial of 804 men with heart failure, investigators compared treatment with enalapril to that with a combination of hydralazine and nitrates.⁶ In the follow-up period, which ranged from 6 months to 5.7 years, 132 of 403 patients (33%) assigned to receive enalapril died, as did 153 of 401 patients (38%) assigned to receive hydralazine and nitrates. The *P* value associated with the difference in mortality is .11.

Looking at this study as an exercise in hypothesis testing (see Part 2B2, “Therapy and Understanding the Results, Hypothesis Testing”) and adopting the usual 5% risk of obtaining a false-positive result, we would conclude that chance cannot be excluded as an explanation of the study results. We would classify this as a negative study (ie, we would conclude that no important difference existed between the treatment and control groups). The investigators also conducted an analysis that compared not only the proportion of patients surviving at the end of the study, but also the time pattern of the deaths occurring in both groups. This survival analysis, which generally is more sensitive than the test of the difference in proportions (see Part 2B2, “Therapy and Understanding the Results, Measures of Association”), showed a nonsignificant *P* value of .08, a result that leads to the same conclusion as the simpler analysis that focused on results at the end of the study. However, the authors also tell us that the *P* value associated with differences in mortality at 2 years (“a point predetermined to be a major endpoint of the trial”) was significant at .016.

At this point, clinicians could be excused for being a little confused. Ask yourself: is this a positive study dictating use of an angiotensin-converting enzyme (ACE) inhibitor instead of the combination of hydralazine and nitrates, or is it a negative study, showing no difference between the two regimens and leaving the choice of drugs open?

SOLVING THE PROBLEM: WHAT ARE CONFIDENCE INTERVALS?

How can clinicians deal with the limitations of hypothesis testing and resolve the confusion? The solution comes from an alternative approach that does not ask about how compatible the results are with the null hypothesis, or whether the P values differ significantly. By contrast, this approach poses two questions: (1) what is the single value most likely to represent the true difference between treatment and control? and (2) given the observed difference between treatment and control, what is the plausible range of differences between them within which the true difference might actually lie? This second question can be answered using *confidence intervals*. Before applying them to resolve the issue of enalapril vs hydralazine and nitrates in patients with heart failure, we will illustrate the use of confidence intervals with a coin-toss experiment.

Suppose that we have a coin that may or may not be balanced. That is, although it may be that the true probability of heads on any individual coin toss is 0.5, it may also be that the true probability is as high as 1.0 in favor of heads (every toss will yield heads) or 1.0 in favor of tails (every toss will yield tails). We now decide to conduct an experiment to determine the true nature of the coin.

We begin by tossing the coin twice, observing one head and one tail. At this point, what is our best estimate of the probability of heads on any given coin toss? Is it the value we have obtained (otherwise known as the *point estimate*), which is 0.5? What is the plausible range within which the true probability of finding a head on any individual coin toss might lie? This range is very wide, and most people would think that the probability might still be as high or higher than 0.9—or as low as or lower than 0.1. In other words, if the true probability of heads on any given coin toss is 0.9, it would still not be terribly surprising if, in any sample of two coin tosses, one were heads and one were tails. Hence, after our two coin tosses we are not much further ahead in determining the true nature of the coin.

We proceed with eight additional coin tosses; after a total of 10 tosses, we have observed five heads and five tails. Our best estimate of the true probability of heads on any given coin toss remains 0.5, the point estimate. The range within which the true probability of heads might plausibly lie has narrowed, however. It is no longer plausible that the true probability of heads is as great as 0.9. That is, if the true probability were 0.9, it would be very unlikely that in a sample of 10 coin tosses, one would observe five tails. People's sense of the range of probabilities that might still be plausible may differ, but most would agree that a probability greater than 0.8 or less than 0.2 is very unlikely.

After 10 coin tosses, values between 0.2 and 0.8 are not all equally plausible. The most likely value for the probability is the point estimate, 0.5, but probabilities close to that point estimate (0.4 or 0.6, for instance) are also quite likely. The further the probability from the point estimate, the less likely it is that the value represents the truth.



Ten coin tosses have still left us with considerable uncertainty about our coin, so we conduct another 40 repetitions. After 50 coin tosses, we have observed 25 heads and 25 tails and our point estimate remains 0.5. We are now beginning to believe that the coin is very unlikely to be extremely biased, and our estimate of the range of probabilities, which is still reasonably consistent with 25 heads in 50 coin tosses, might be 0.35 to 0.65. This range still is quite wide and we may persist with another 50 repetitions. If after 100 tosses we observed 50 heads, we might guess that the true probability is unlikely to be more extreme than 0.40 or 0.60. If we were willing to endure the tedium of 1000 coin tosses and if we observed 500 heads, we would be very confident (but still not certain) that our coin is minimally, if at all, biased.

What we have done through this experiment is to use common sense to generate confidence intervals around an observed proportion, 0.5. In each case, the confidence interval represents the range within which the truth plausibly lies. The smaller the sample size, the wider the confidence interval. As the sample size gets very large, we become increasingly certain that the truth is not far from the point estimate we have calculated from our experiment and the confidence interval is smaller.

It is fortunate that, since people's common sense differs considerably, we can turn to statistical techniques for precise estimation of confidence intervals. To use these techniques, we must first be a little more specific about what we mean by "plausible." In our coin-toss example, we might ask "what is the range of probabilities within which, 95% of the time, the truth would lie?" Table 2B2-2 presents the actual 95% confidence intervals around the observed proportion of 0.5 for our experiment. If we need not be quite so certain, we could ask about the range within which the true value would lie 90% of the time. This 90% confidence interval, also presented in Table 2B2-2, is somewhat narrower.

TABLE 2B2-2

Confidence Intervals Around a Proportion of 0.5 in a Coin-Toss Experiment

Number of Coin Tosses	Observed Result	95% Confidence Interval	90% Confidence Interval
2	1 head, 1 tail	0.01–0.99	0.03–0.98
10	5 heads, 5 tails	0.19–0.81	0.22–0.78
50	25 heads, 25 tails	0.36–0.65	0.38–0.62
100	50 heads, 50 tails	0.40–0.60	0.41–0.59
1000	500 heads, 500 tails	0.47–0.53	0.47–0.53

The coin-toss example also illustrates how the confidence interval tells you whether the study is large enough to answer the research question. If you wanted to be reasonably sure that the bias was no greater than 10% (that is, the ends of the confidence interval are within 10% of the point estimate), you would need approximately 100 coin tosses. If you needed greater precision—with 3% in either direction—1000 coin tosses would be required. All you have to do to obtain greater precision is to make more measurements. In clinical research, this involves enrolling more patients or increasing the number of measurements in each patient who is enrolled.

USING CONFIDENCE INTERVALS TO INTERPRET THE RESULTS OF CLINICAL TRIALS

How do confidence intervals help us interpret the results of the trial of vasodilators in patients with heart failure? The mortality in the ACE inhibitor arm was 33% and in the hydralazine plus nitrate group it was 38%, an absolute difference of 5%. The difference of 5% is the point estimate, our best single estimate of the mortality benefit from using an ACE inhibitor. The 95% confidence interval around this difference works out to -1.2% to 12% .

How can we now interpret the study results? The most likely value for the mortality difference between the two vasodilator regimens is 5%, but the true difference may be as high as 1.2% in favor of the combination of hydralazine and nitrates or as high as 12% in favor of the ACE inhibitor. Values progressively farther from 5% will be less and less probable. We can conclude that patients offered ACE inhibitors will most likely (but not certainly) die later than patients offered hydralazine and nitrates—but the magnitude of the difference may be either trivial or quite large. This way of understanding the results avoids the yes/no dichotomy of hypothesis testing and the possible consequences of spending time and energy deciding about the legitimacy of the authors' focus on mortality at 2 years. It also obviates the need to argue whether the study should be considered positive or negative. One can conclude that, all else being equal, an ACE inhibitor is the appropriate choice for patients with heart failure, but the strength of this inference is weak. Toxicity, expense, and evidence from other studies would all bear on the final treatment decision (see Part 1F, "Moving From Evidence to Action"). Since a number of large randomized trials have now shown a mortality benefit from ACE inhibitors in patients with heart failure,⁷ one can confidently recommend this class of agents as the treatment of choice.



INTERPRETING APPARENTLY “NEGATIVE” TRIALS

Another example of the use of confidence intervals in interpreting study results comes from the results of the Swedish Co-operative Stroke Study, a randomized trial that was designed to determine whether patients with cerebral infarction might have fewer subsequent strokes if they took aspirin.^{8,9} The investigators gave placebos to 252 patients, of whom 18 (7%) subsequently had nonfatal stroke. They also gave aspirin to 253 patients, of whom 23 (9%) had recurrent nonfatal stroke. The point estimate from these results is a 2% increase in the incidence of strokes among those patients in the aspirin group.

This trial of more than 500 patients might appear to exclude any possible benefit from aspirin. The 95% confidence interval on the absolute difference of 2% in favor of placebo, however, is from 7% in favor of placebo to 3% in favor of aspirin. Were the truth that 3% of the patients who would otherwise have strokes been spared had they taken aspirin, many patients would want to receive that drug. This would represent a 43% relative risk reduction, suggesting that we would need to treat only 33 patients to prevent a stroke. One can thus conclude that the trial has not excluded a patient-important benefit and, in that sense, was not large enough.

This example emphasizes that many patients must participate if trials are to generate precise estimates of treatment effects. In addition, it illustrates why we recommend that, whenever possible, clinicians turn to systematic reviews that pool data from the most valid studies.¹⁰ In this case, such an overview shows that administration of antiplatelet agents in patients with transient ischemic attack or stroke reduces the relative risk of subsequent events by approximately 25% (with confidence intervals ranging from approximately 19% to 31%).¹¹ Given these data, many patients whose event rates without treatment would be over 10% (a number needed to treat of 50 or less) or even 5% (a number needed to treat of 100 or less) would be enthusiastic about taking aspirin.

This example also illustrates that when you see an apparently negative trial (one that, in our previous hypothesis-testing framework, fails to exclude the null hypothesis), you can focus on the upper end of the confidence interval (that is, the end that suggests the largest benefit from treatment). If the upper boundary of the confidence interval excludes any important benefit of treatment, you can conclude the trial is definitively negative. If, on the other hand, the confidence interval includes an important benefit, the possibility has not been ruled out that the treatment still might be worthwhile.

This logic of the negative trial is crucial in the interpretation of studies designed to help determine whether we should substitute a treatment that is less expensive, easier to administer, or less toxic for an existing treatment. In such an *equivalence study*, we will be ready to make the substitution only if we are sure that the standard treatment does not have important additional benefit beyond the less expensive or more convenient substitute. We will be confident that we have excluded the possibility of important additional benefit of the standard treatment if the upper boundary of the confidence interval around the difference is below our threshold.

INTERPRETING APPARENTLY “POSITIVE” TRIALS

How can confidence intervals be informative in a positive trial (one that, in the previous hypothesis-testing framework, makes chance an unlikely explanation for observed differences between treatments)? In another double-blind randomized controlled trial of patients with heart failure, treatment with enalapril was compared to that with placebo.¹² Of 1285 patients randomized to the ACE inhibitor, 613 (48%) died or were hospitalized for accelerated heart failure, whereas 736 (57%) of 1284 patients in the placebo group experienced one of these adverse outcomes. The point estimate of the difference in death or hospitalization for heart failure is 10%, and the 95% confidence interval is 6% to 14%. Thus, the smallest effect of the ACE inhibitor that is compatible with the data is a 6% reduction in the number of patients with the adverse outcomes. If you consider it worthwhile to treat 17 patients to prevent one patient from dying or developing heart failure (6% is equivalent to about one in 17), then this represents a definitive trial. If, before treating, you would require a greater reduction than 6% in the proportion of patients who are spared an adverse event, a larger trial (with a correspondingly narrower confidence interval) would be required.

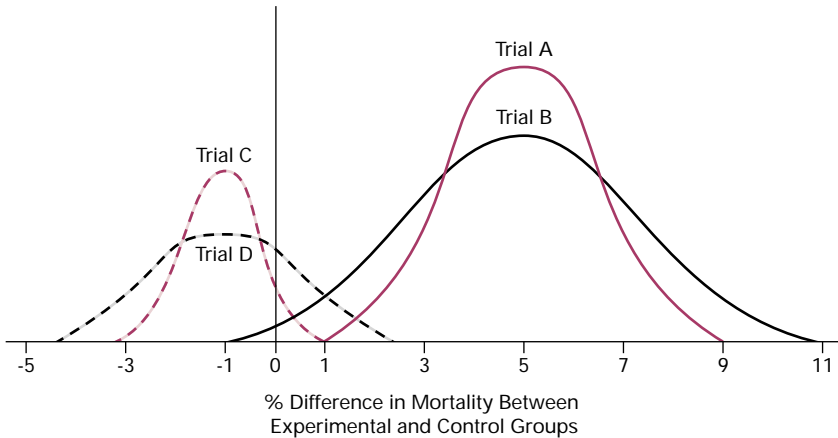
WAS THE TRIAL LARGE ENOUGH?

As implied in our discussion to this point, confidence intervals provide a way of answering the question: “Was the trial large enough?” We illustrate the approach in Figure 2B2-2. In this figure, we present the distribution of randomized trial results you would expect from two treatments—one that results in an absolute reduction in mortality of 5% and one that results in an absolute increase in mortality of 1%. The vertical line in the center of the figure represents an absolute risk reduction of zero, when the experimental and control groups have exactly the same mortality. Values to the right of the vertical line represent results in which the treated group had a lower mortality than the control group. Values to the left of the vertical line represent results in which the treated group fared worse and had a higher mortality rate than the control group.



FIGURE 2B2-2

Deciding Whether a Trial Is Definitive: Distributions of the Results of Trials of Two Therapies



A represents the results of large trials of a therapy with an absolute mortality reduction of 5%; B represents the results of smaller trials of a therapy with an absolute reduction in mortality of 5%; C represents the results of large trials of a therapy with an absolute mortality increase of 1%; D represents the results of smaller trials of a therapy with an absolute reduction in mortality by 1%.

Reproduced with permission from the Canadian Medical Association.

For each of the two treatments, we present two distributions of results: one for a set of trials with a relatively small sample size, and one for a set of trials with a relatively large sample size. For each of the four distributions, the highest point of the distribution represents the underlying truth, the actual change in mortality. Distributions A and B come from the trials of the therapy that reduced mortality by 5%, and distributions C and D come from trials of the therapy that increased mortality by 1%.

Now, suppose we assume that absolute reductions in mortality greater than 1% warrant treatment. That is, the benefits outweigh the risks and costs whenever the absolute reduction in risk is 1% or greater (see Part 1F, “Moving From Evidence to Action”; see also Part 2F, “Moving From Evidence to Action, Grading Recommendations—A Quantitative Approach”), whereas reductions less than 1% do not warrant treatment (that is, the risks outweigh the benefits). For instance, if experimental treatment results in a true reduction in mortality from 5% to less than 4%, we would want to use the treatment. If, on the other hand, the true reduction in mortality was 5% to 4.5%, we would consider that the experimental treatment was not worth the associated toxicity and expense. What implications does this have for the way we will interpret the results of studies of this treatment?

In distribution A, more than 95% of the distribution lies above an absolute risk reduction of 1% (distribution A, like the others, depicts a simplified presentation of the situation—probabilities never actually sink to zero). Based on trials of

this therapy and on this sample size, 95% confidence intervals would, in most instances, exclude an absolute risk reduction as small as 1%. In such trials, we could be confident that the true treatment effect is above our threshold, 1%, and we have a definitive positive trial. That is, we would be very confident that the true reduction in risk is greater than 1% (and, most likely, is appreciably greater), suggesting that many patients would be interested in receiving the treatment. The sample size in such trials would be adequate to demonstrate that the treatment provides a clinically important benefit.

Distribution B also comes from trials of a therapy that reduces mortality by 5%, but these trials include fewer patients. Whereas some of these trials would exclude the null hypothesis (that is, no difference is assumed between the treatment and control groups), many of the 95% confidence intervals would include mortality reductions less than 1%. When the 95% confidence interval includes values less than 1%, the data are consistent with an absolute risk reduction less than 1%. For such trials, we are left in doubt that the treatment effect is really greater than our threshold. Such trials would still be perceived as positive, but their results would not be definitive. The sample size of these trials would be inadequate to definitively establish the appropriateness of administering the experimental treatment.

Distribution C shows the results of a set of trials, all of which would be negative in that they would not exclude the null hypothesis of “no treatment effect.” On average, investigators conducting these trials would observe a mortality rate that was 1% higher in the treatment group than in the control group. Most such trials would generate a narrow 95% confidence interval, all of which would lie to the left of our 1% threshold. The fact that the upper limit of the confidence interval is less than 1% would mean that we can be very confident that, if there is a benefit, it is very small and is unlikely to be appreciably greater than the risks, costs, and inconvenience of therapy. These trials would therefore exclude any patient-important benefit of treatment and they could be considered definitive. We would therefore dismiss the experimental treatment—at least for this type of population.

Distribution D comes from the same therapy as is reflected in distribution C, in which the mortality is 1% higher in the experimental group than in the control group. Distribution D, however, depicts trials with smaller sample size and, consequently, a much wider distribution of results. Because the confidence interval of most of these trials would include an appreciable portion that lies above our 1% threshold, we would conclude that it remains plausible (though unlikely) that the true effect of the experimental treatment is a reduction in mortality greater than 1%. Although we would still refrain from using this treatment (indeed, we would conclude it most likely kills people), we would not totally dismiss it. Most trials from distribution D, therefore, would not be definitive, and we would require larger trials enrolling more patients to exclude a clinically



important treatment effect.

CONCLUSION

We can restate our message as follows: in a positive trial establishing that the effect of treatment is greater than zero, look to the lower boundary of the confidence interval to determine whether sample size has been adequate. If this lower boundary—the smallest plausible treatment effect compatible with the data—is greater than the smallest difference that you consider important, the sample size is adequate and the trial is definitive. If the lower boundary is less than this smallest important difference, the trial is nondefinitive and further trials are required.

In a negative trial, look to the upper boundary of the confidence interval to determine whether sample size has been adequate. If this upper boundary, the largest treatment effect compatible with the data, is less than the smallest difference that you consider important, the sample size is adequate and the trial is definitively negative. If the upper boundary exceeds the smallest important difference, there may still be an important positive treatment effect, the trial is nondefinitive, and further trials are required.

References

1. Simon R. Confidence intervals for reporting results of clinical trials. *Ann Intern Med.* 1986;105:429-435.
2. Gardner MJ, Altman DG, eds. *Statistics With Confidence: Confidence Intervals and Statistical Guidelines.* London: BMJ Publishing Group; 1989.
3. Bulpitt CJ. Confidence intervals. *Lancet.* 1987;1:494-497.
4. Pocock SJ, Hughes MD. Estimation issues in clinical trials and overviews. *Stat Med.* 1990;9:657-671.
5. Braitman LE. Confidence intervals assess both clinical significance and statistical significance. *Ann Intern Med.* 1991;114:515-517.
6. Cohn JN, Johnson G, Ziesche S, et al. A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure. *N Engl J Med.* 1991;325:303-310.
7. Garg R, Yusuf S. Overview of randomized trials of angiotensin-converting enzyme inhibitors on mortality and morbidity in patients with heart failure. Collaborative Group on ACE Inhibitor Trials. *JAMA.* 1995;273:1450-1456.
8. Britton M, Helmers C, Samuelsson K. High-dose salicylic acid after cerebral infarction: a Swedish co-operative study. *Stroke.* 1997;18:325.

9. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical Epidemiology; A Basic Science for Clinical Medicine*. Boston: Little, Brown and Company; 1991:218-220.
10. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. *CMAJ*. 1988;138:697-703.
11. Antiplatelet Trialists' Collaboration. Secondary prevention of vascular disease by prolonged antiplatelet treatment. *BMJ*. 1988;296:320-331.
12. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *N Engl J Med*. 1991;325:293-302.