Basic statistics for clinicians: 2. Interpreting study results: confidence intervals

Gordon Guyatt,*† MD; Roman Jaeschke,*† MD; Nancy Heddle,‡ MSc; Deborah Cook,*† MD; Harry Shannon,* PhD; Stephen Walter,* PhD

Abstract • Résumé

In the second of four articles, the authors discuss the "estimation" approach to interpreting study results. Whereas, in hypothesis testing, study results lead the reader to reject or accept a null hypothesis, in estimation the reader can assess whether a result is strong or weak, definitive or not. A confidence interval, based on the observed result and the size of the sample, is calculated. It provides a range of probabilities within which the true probability would lie 95% or 90% of the time, depending on the precision desired. It also provides a way of determining whether the sample is large enough to make the trial definitive. If the lower boundary of a confidence interval is above the threshold considered clinically significant, then the trial is positive and definitive; if the lower boundary is somewhat below the threshold, the trial is positive, but studies with larger samples are needed. Similarly, if the upper boundary of a confidence interval is below the threshold considered significant, the trial is negative and definitive. However, a negative result with a confidence interval that crosses the threshold means that trials with larger samples are needed to make a definitive determination of clinical importance.

Dans le deuxième article d'une série de quatre, les auteurs discutent de la façon «estimative» d'interpréter les résultats des études. Même si, dans un test d'hypothèse, les résultats de l'étude mènent le lecteur à rejeter ou à accepter une hypothèse nulle, dans une estimation, le lecteur peut évaluer si un résultat est fort ou faible, concluant ou non. On calcule un intervalle de confiance d'après les résultats observés et la taille de l'échantillon. Cet intervalle fournit une gamme de probabilités qui comprendrait la probabilité réelle dans 95 % ou 90 % du temps, selon le degré de précision désiré. Il fournit également une façon de déterminer si la taille de l'échantillon est assez grande pour que l'essai soit concluant. Si la limite inférieure d'un intervalle de confiance se situe au-dessus du seuil considéré comme significatif du point de vue clinique, l'essai est alors positif et concluant; si la limite inférieure se trouve quelque peu en-dessous du seuil, l'essai est positif, mais il faut effectuer d'autres études avec des échantillons de plus grande taille. De même, si la limite supérieure d'un intervalle de confiance se trouve au-dessous du seuil considéré comme significatif, l'essai est négatif et concluant. Toutefois, un résultat négatif obtenu avec un intervalle de confiance dont les valeurs recoupent le seuil signifie qu'il faut procéder à d'autres essais avec des échantillons de taille plus grande pour obtenir une détermination définitive ayant une importance clinique.

In our first article in this series we explained hypothesis testing, which involves estimating the likelihood that observed results of an experiment would have occurred by chance if a null hypothesis — that there was no difference between the effects of a treatment and a control condition — were true. The limitations of hypothesis testing have been increasingly recognized, and an alternative approach, called estimation, is becoming more popular. Several authors[1–5] have outlined the concepts that we will introduce in this article, and their discussions may be read to supplement our explanation.

An example from our first article illustrates the limitations of the hypothesis-testing approach. In the results of this trial, the decision to reject the null hypothesis rests on the analysis one prefers.

Interpreting study results: How should we treat heart failure?

In a double-blind randomized trial, treatment with enalapril was compared with therapy with a combination of

*From the departments of *Clinical Epidemiology and Biostatistics, †Medicine and ‡Pathology, McMaster University, Hamilton, Ont.*

Dr. Cook is a recipient of a Career Scientist Award from the Ontario Ministry of Health. Dr. Walter is the recipient of a National Health Scientist Award from Health Canada.

***Reprint requests to:** Dr. Gordon Guyatt, Rm. 2C12, McMaster University Health Sciences Centre, 1200 Main St. W, Hamilton ON L8N 3Z5*

This is the second article in a series of four, to appear in the January and February 1995 issues of CMAJ.

← For prescribing information see page 289

hydralazine and nitrates in 804 men with congestive heart failure.[6] During the period patients were followed up, from 6 months to 5.7 years, 33% (132/403) of the patients assigned to enalapril died, as did 38% (153/401) of those assigned to hydralazine and nitrates. The p value associated with the difference in mortality, determined by a χ^2 test, was 0.11.

If one considered this study an exercise in hypothesis testing and adopted the usual threshold for Type I error of $p = 0.05$, one would conclude that chance is an adequate explanation for the study results. One would classify this as a "negative" study, i.e., a study showing no important difference between treatment and control groups. However, the investigators also used their data to conduct a "survival analysis," which is generally more sensitive than a test of the difference in proportions. The p value for mortality obtained from the survival analysis was 0.08, a result that leads to the same conclusion as the simpler χ^2 test. However, the authors also reported that the p value associated with differences in mortality after 2 years ("a point predetermined to be a major end point of the trial") was 0.016.

The reader could be excused for experiencing a little confusion. Do these results mean that this is a "positive" study supporting the use of an angiotensin-converting-enzyme (ACE) inhibitor (enalapril) rather than the combination of hydralazine and nitrates or a "negative" study leaving open the choice of drug treatments?

Solving the problem: confidence intervals

How can the limitations of hypothesis testing be remedied and the confusion resolved? The solution is found in an alternative approach that does not determine the compatibility of the results with the null hypothesis. This approach poses two questions: What is the single value most likely to represent the true difference between the treatment and control groups? and, given the observed difference between treatment and control groups, What is the plausible range of differences within which the true difference may lie? The second question can be answered with the use of confidence intervals. Before applying confidence intervals to resolve the issue of the benefits of enalapril versus those of hydralazine and nitrates, we will illustrate the use of confidence intervals with a coin-toss experiment similar to the one we conducted in the first article.

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

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

We proceed with another eight coin tosses; after a total of 10, 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 may plausibly lie has, however, narrowed. It is no longer plausible that the true probability of heads is as great as 0.9; with such a high probability, it would be very unlikely that one would observe 5 tails in a sample of 10 coin tosses. People's sense of the range of plausible probabilities may differ, but most would agree that a probability greater than 0.8 or less than 0.2 is very unlikely.

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

Ten tosses have still left us with considerable uncertainty about our coin, and 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 now believe that the coin is very unlikely to be extremely biased, and our estimate of the range of probabilities that is reasonably consistent with 25 heads in 50 coin tosses is 0.35 to 0.65. This is still a wide range, and we may persist with another 50 repetitions. If after 100 tosses we had 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 we observed 500 heads, we would be very confident (but still not certain) that our coin is minimally, if at all, biased.

In this experiment we have used 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, the wider the confidence interval. As the sample becomes larger, we are increasingly certain that the truth is not far from the point estimate we have observed from our experiment.

Since people's "common-sense" estimate of the plausible range differs considerably, we can turn to statistical techniques for precise estimation of confidence intervals. To use these techniques we must be more specific about what we mean by "plausible." In our coin toss example we could ask What is the range of probabilities within which, 95% of the time, the true probability would lie? The actual 95% confidence intervals around the observed proportion of 0.5 for our coin toss experiment are given in Table 1. If we do not need to be 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 1, is somewhat narrower.

The coin toss example also illustrates how the confidence interval tells us whether the sample is large enough to answer the research question. If you wanted to be reasonably sure that any bias in the coin is no greater than 10% (that is, the confidence interval is within 10% of the point estimate) you would need approximately 100 coin tosses. If you needed greater precision — with 3% of the point estimate — 1000 coin tosses would be required. To obtain greater precision all you must do is make more measurements. In clinical research, this involves enrolling more subjects or increasing the number of measurements in each subject enrolled. (But take care: increasing precision by enlarging the sample or increasing the number of measurements does not compensate for poor study design.[7–9])

Using confidence intervals to interpret study results

How can confidence intervals help us interpret the results of the trial to determine different effects of vasodilators in the treatment of heart failure? In the ACE-inhibitor arm of the trial 33% of the patients died, and in the group assigned to hydralazine and nitrates 38% died, yielding an absolute difference of 5%. This difference is the point estimate, our best single estimate of the benefit in lives saved from the use of an ACE inhibitor. The 95% confidence interval around this difference is −1.2% to 12%.

How can we now interpret the study results? The most likely value for the difference in mortality between the two vasodilator regimens is 5%, but the true difference may be up to 1.2% in favour of hydralazine and nitrates or up to 12% in favour of the ACE inhibitor. Values farther from 5% are less and less probable. We can conclude that patients offered ACE inhibitors most likely (but not certainly) will die later than patients offered hydralazine and nitrates; however, the magnitude of the difference in expected survival may be trivial or large. This way of understanding the results avoids the Yes–No dichotomy that results from hypothesis testing, the expenditure of time and energy to evaluate the legitimacy of the authors' end point of mortality after 2 years, and consideration of whether the study is "positive" or "negative" on the basis of the results. One can conclude that, all else being equal, an ACE inhibitor is the appropriate choice for patients with heart failure, but that the strength of this inference is weak. The toxic effects and cost of the drugs, and evidence from other studies, would all bear on the treatment decision. Since several large randomized trials have now shown that a benefit is gained from the use of ACE inhibitors in patients with heart failure,[10,11] one can confidently recommend this class of agents as the treatment of choice.

Interpreting trials that appear to be "negative"

In another example of the use of confidence intervals in interpreting study results, Sackett and associates[12] examined results from the Swedish Co-operative Stroke Study, a trial designed to determine whether patients with cerebral infarcts would have fewer subsequent strokes if they took acetylsalicylic acid (ASA).[13] The investigators gave placebos to 252 patients in the control group, of whom 7% (18) had a subsequent nonfatal stroke, and ASA to 253 patients in the experimental group, of whom 9% (23) had a nonfatal stroke. The point estimate was therefore a 2% increase in strokes with ASA prophylaxis. The results certainly did not favour the use of ASA for prevention of stroke.

The results of this large trial, involving more than 500 patients, may appear to exclude any possible benefit from ASA. However, the 95% confidence interval around the point estimate of 2% in favour of placebo is from 7% in favour of placebo to 3% in favour of ASA. If, in fact, 3% of patients who had strokes would have been spared if they had taken ASA, one would certainly want to administer the drug. By treating 33 patients, one stroke could be prevented. Thus, one can conclude that the Swedish study did not exclude a clinically important benefit and, in that sense, did not have a large enough sample.

As this example emphasizes, many subjects are needed in order to generate precise estimates of treatment effects; this is why clinicians are turning more and more to rigorous meta-analyses that pool data from the most valid studies.[14]

Table I: 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

In the case of ASA prophylaxis for recurrent stroke, such a meta-analysis showed that antiplatelet agents given to patients with a previous transient ischemic attack (TIA) or stroke reduced the risk of a subsequent TIA or stroke by approximately 25% (confidence interval approximately 19% to 31%). This benefit is great enough that most clinicians will want to treat such patients with ASA.[15]

This example also illustrates that, when one sees results of an apparently "negative" trial (one that, in a hypothesis-testing framework, would fail to exclude the null hypothesis), one should pay particular attention to the upper end of the confidence interval, that is, the end that suggests the largest benefit from treatment. If even the smallest benefit of clinical importance lies above the upper boundary of the confidence interval, the trial is definitively negative. In contrast, if clinically important benefits fall within the confidence interval, the trial has not ruled out the possibility that the treatment is worth while.

INTERPRETING TRIALS THAT APPEAR TO BE "POSITIVE"

How can confidence intervals provide information about the results of a "positive" trial — results that, in the previous hypothesis-testing framework, would be definitive enough to exclude chance as the explanation for differences between results of treatments? In another double-blind randomized trial of treatments for heart failure, the effect of enalapril was compared with that of a placebo.[11] Of 1285 patients randomly assigned to receive the ACE inhibitor, 48% (613) died or were admitted to hospital for worsening heart failure, whereas 57% (736/1284) of patients who received placebo died or required hospital care. The point estimate of the difference in death or hospital admission for heart failure was 10%, and the 95% confidence interval was 6% to 14%. Thus, the smallest true effect of the ACE inhibitor that is compatible with the data is a 6% (or about 1 in 17) reduction in the number of patients with these adverse outcomes. If it is considered worth while to treat 17 patients in order to prevent one death or heart failure, this trial is definitive. If, before using a drug, you require a reduction of more than 6% in the proportion of patients who are spared death or heart failure, a larger trial (with a correspondingly narrower confidence interval) would be required.

WAS THE TRIAL LARGE ENOUGH?

Confidence intervals provide a way of answering the question Was the trial large enough? We illustrate this approach in Fig. 1. Each of the distribution curves represents the results of one hypothetical randomized trial of an experimental treatment to reduce mortality (trials A, B, C and D). The vertical line at 0% represents a risk reduction of 0: a result at this value means that mortality in the experimental and control groups is exactly the same. Values to the right of the vertical line represent results in which the experimental group had a lower mortality than the control group; to the left of the vertical line, results in which the experimental group fared worse, with a higher mortality than the control group.

The highest point of each distribution represents the result actually observed (the point estimate). In trials A and B the investigators observed that mortality was 5% lower in the experimental group than in the control group. In trials C and D they observed that mortality was 1% higher in the experimental group than in the control group.

The distributions of the likelihood of possible true results of each trial are based on the point estimate and the size of the sample. The point estimate is the single value that is most likely to represent the true effect. As you can see, values farther from the results observed are less likely than values closer to the point estimate to represent the true difference in mortality.

Now, suppose we assume that an absolute reduction in mortality of greater than 1% means that treatment is warranted (that is, such a result is clinically important), and a reduction of less than 1% means that treatment is not warranted (that is, the result is trivial). For example, if the experimental treatment results in a true reduction in mortality from 5% to 4% or less, we would want to use the treatment. If, on the other hand, the true reduction in mortality was from 5% to 4.5%, we would consider the benefit of the experimental treatment not to be worth the associated toxic effects and cost. What are the implications of this decision for the interpretation of the results of the four studies?

In trial A the entire distribution and, hence, the entire 95% confidence interval lies above the threshold risk reduction of 1%. We can therefore be confident that the true treatment effect is above our threshold, and we have a definitive "positive" trial. That is, we can be very confident that the true reduction in risk is greater — probably appreciably greater — than 1%; this leaves little doubt that we should administer the treatment to our patients. The sample size in this trial was adequate to show that the treatment provides a clinically important benefit.

Trial B has the same point estimate of treatment effect as trial A (5%) and is also "positive" ($p < 0.05$). In a hypothesis

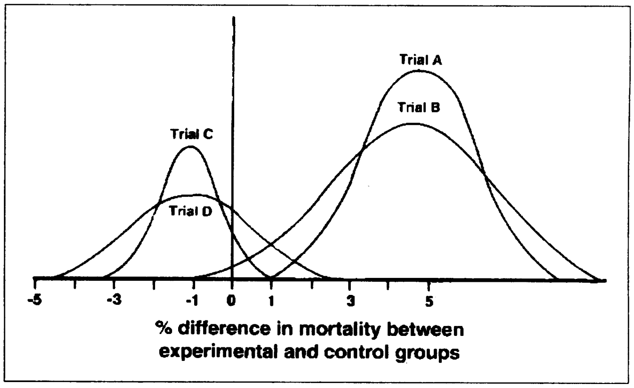

Fig. 1: Distributions of the likelihood of the true results of four trials (A, B, C and D). Trial A is a definitive and trial B a nondefinitive positive trial. Trial C is a definitive and trial D a nondefinitive negative trial.

test, the null hypothesis would be rejected. However, more than 2.5% of the distribution is to the left of the 1% threshold. In other words, the 95% confidence interval includes values less than 1%. This means that the data are consistent with an absolute risk reduction of less than 1%, so we are left with some doubt that the treatment effect is really greater than our threshold. This trial is still "positive," but its results are not definitive. The sample in this trial was inadequate to establish definitively the appropriateness of administering the experimental treatment.

Trial C is "negative"; its results would not lead to the rejection of the null hypothesis in a hypothesis test. The investigators observed mortality 1% higher in the treatment than in the control group. The entire distribution and, therefore, the 95% confidence interval lie to the left of our 1% threshold. Because the upper limit of the distribution is 1%, we can be very confident that, if there is a positive effect, it is trivial. The trial has excluded any clinically important benefit of treatment, and it can be considered definitive. We can therefore decide against the use of the experimental treatment, at least for this type of patient.

The result of trial D shows the same difference in absolute risk as that of trial C: mortality 1% higher in the experimental than in the control group. However, trial D had a smaller sample and, as a result, a much wider distribution and confidence interval. Since an appreciable portion of the confidence interval lies to the right of our 1% threshold, it is plausible (although unlikely) that the true effect of the experimental treatment is a reduction in mortality of greater than 1%. Although we would refrain from using this treatment (indeed, the most likely conclusion is that it kills people), we cannot completely dismiss it. Trial D was not definitive, and a trial involving more patients is required to exclude a clinically important treatment effect.

Conclusions

We can restate our interpretation of confidence intervals as follows. In a "positive" trial — one that establishes that the effect of treatment is greater than zero — look at the lower boundary of the confidence interval to determine whether the size of the sample is adequate. The lower boundary represents the smallest plausible treatment effect compatible with the data. If it is greater than the smallest difference that is clinically important, the sample size is adequate and the trial definitive. However, if it is less than this smallest important difference, the trial is not definitive and further trials are required. In a "negative" trial — the results of which do not exclude the possibility that the treatment has no effect — look at the upper boundary of the confidence interval to determine whether the size of the sample is adequate. If the upper boundary — the largest treatment effect compatible with the data — is less than the smallest difference that is clinically important, the size of the sample is adequate, and the trial is definitively negative. If the upper boundary exceeds the smallest difference considered important, there may be an important positive treatment effect, the trial is not definitive, and further trials are required.

In this discussion we have examined absolute differences in proportions of patients who died while receiving two different treatments. In the next article in this series, we will explain how to interpret other ways investigators present treatment effects, including odds ratios and relative risk.

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*, British Medical Journal, London, England, 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. Oxman AD, Sackett DL, Guyatt GH and the Evidence-Based Medicine Working Group: Users' guides to the medical literature: I. How to get started. *JAMA* 1993; 270: 2093–2095
8. Guyatt GH, Sackett DL, Cook DJ and the Evidence-Based Working Group: Users' guides to the medical literature: II. How to use an article about therapy or prevention. A. Are the results of the study valid? *JAMA* 1993; 270: 2598–2601
9. Guyatt GH, Sackett DL, Cook DJ and the Evidence-Based Working Group: Users' guides to the medical literature: II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA* 1994; 271: 59–63
10. Mulrow CD, Mulrow JP, Linn WD et al: Relative efficacy of vasodilator therapy in chronic congestive heart failure. *JAMA* 1988; 259: 3422–3426
11. 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
12. Sackett DL, Haynes RB, Guyatt GH et al: *Clinical Epidemiology, a Basic Science for Clinical Medicine*, Little, Brown and Company, Boston, 1991: 218–220
13. Britton M, Helmers C, Samuelsson K: High-dose salicylic acid after cerebral infarction: a Swedish co-operative study. *Stroke* 1987; 18: 325
14. Oxman AD, Cook DJ, Guyatt GH and the Evidence-Based Medicine Working Group: Users' guides to the medical literature: VI. How to use an overview. *JAMA* 1994; 272: 1367–1371
15. Antiplatelet trialists' collaboration: Secondary prevention of vascular disease by prolonged antiplatelet treatment. *BMJ* 1988; 296: 320–331