FROM RESEARCH TO PRACTICE

A Clinician's View of Statistics

Carl D. Atkins, MD

Clinical trials provide a major source of the new knowledge that underlies medical progress. Statistics are essential for interpreting these studies, but clinicians have little guidance in their use. In this article, I will show how to use statistics to help incorporate the findings of clinical trials into daily practice. I will do this with the nonresearcher in mind, without the use of mathematical formulas or technical jargon.

Let us start with a clinical example. A 62-year-old man with multiple myeloma is in remission after chemotherapy. His medically knowledgeable friend has told him that all patients with multiple myeloma should receive treatment with interferon. However, his own reading reveals that this treatment is expensive and often has side effects. The patient turns to you for guidance. "How important is this treatment?" he asks. You search the literature and find three good, randomized trials comparing treatment with interferon to no maintenance treatment for patients with multiple myeloma.¹⁻³ We will examine how to interpret the statistics of these trials, shown in Table 1, to help answer this patient's question.

Table 1. Comparison of Survival Rates for Patients with Multiple Myeloma in Remission Treated With or Without Interferon

Study	Three-Year Survival*			
	Interferon	No Interferon	p Value	95% CI for Difference
Mandelli et al. ¹	0.75	0.55	.04	.02 to .38
Peest et al. ²	0.61	0.71	.31	-.27 to +.07
Westin et al. ³	0.61	0.55	.50	-.11 to +.23

*Survival values were obtained from survival curves as printed in figures, and confidence intervals and p values were calculated from the data. All patients were assumed to have been followed for a minimum of 3 years.

From Research to Practice, a Journal series, presents articles to heighten the clinician's awareness of research and methodology issues that have direct relevance to practice. If you wish to submit a manuscript for consideration for this series, please contact Cynthia D. Mulrow, MD, MSc, Associate Editor, at mulrow@uthscsa.edu, or contact the Journal of General Internal Medicine at (215) 823-4471 to receive the appropriate guidelines.

P VALUES DO NOT TELL THE WHOLE STORY

Table 1 shows both *p* values and confidence intervals for each trial. In evaluating clinical trials, I am going to recommend that you ignore the *p* values whenever you can and rely instead on confidence intervals to interpret the results. To see why, let us look at the difference between what *p* values and confidence intervals tell us.

What is a *p* value? The *p* value gives the probability of obtaining results at least as disparate as those observed by repeatedly choosing samples from the same population of patients, i.e., when there is no underlying difference between the two samples. The *p* value thus reflects the differences between two groups that can be expected from random sampling variation, or chance. The closer the *p* value is to 0, the less the difference is likely to be *completely* due to chance. The usual characterization of clinical trials as positive if *p* < .05 stems from a desire to avoid concluding that a treatment is effective unless one can be sure that the results are not completely due to chance.⁴ However, when *p* < .05, it does not mean that the difference observed is *true*, it just means that the difference between groups is unlikely to be 0 (less than a 5% chance). By the same token, when *p* > .05, it does not mean that there is *no* difference between groups, it just means that the observed difference *could* occur by chance alone more than 5% of the time.

If we make treatment decisions based on *p* values, we will be recommending treatment if we are 95% or more sure that there is some benefit of treatment, *regardless of how small that benefit might be*. Similarly, we will not recommend treatment if we are only 94% or less sure that the treatment is any better, *regardless of how large the difference might be*. However, this is *not* what our patient asked us to tell him. What he needs is some *estimate* of the difference due to treatment, with some measure of the *precision* of that estimate.

This is where the confidence interval comes in. This interval defines the range of values that are consistent with the difference measured.⁵ Values outside this range

Received from the South Shore Hematology-Oncology Associates, Rockville Centre, NY.

Address correspondence and reprint requests to Dr. Atkins: South Shore Hematology-Oncology Associates, 242 Merrick Rd., Suite 301, Rockville Centre, NY 11570-5254.

would be expected to give study results less than 5% of the time. Thus, if we measure a 10% difference between groups, the true difference could be somewhat higher or lower. If the 95% confidence interval is 5% to 15%, there is a 95% likelihood that the observed result was generated by a true treatment difference of 5% to 15%. Table 1 illustrates how confidence intervals help us interpret our studies better than p values do.

In the next-to-last column of Table 1, we see that the p values for three studies are very different. Peest et al.² and Westin et al.³ interpreted their results as showing that interferon did not improve survival because in their studies $p > .05$. However, Mandelli's result of $p = .04$ led him to recommend interferon treatment for all patients with multiple myeloma in remission. The p values are conflicting, and they provide no means to resolve that conflict.

Now look at the confidence intervals in the last column in Table 1. We can see that the confidence intervals for all three studies overlap. This means that the studies which do not demonstrate a statistically significant result do not necessarily contradict the study which does. They show that the measured effects *might* be due to chance, not that they *are* due to chance. The *different* results are still *consistent* with one another. If interferon treatment actually results in a 5% improvement in 3-year survival, the differences among the three study results could simply be due to random sampling variation. The problem with the p values is that they did not tell us how different the results could have been without generating a statistically significant result.⁶ Some authors choose to address this issue by estimating the power of the study, but this is superfluous (Goodman and Berlin offer an excellent discussion of the merits of confidence intervals over power calculations⁷). Instead, the confidence intervals draw attention to the problem—because all three trials are small, the estimates of effect are imprecise. One appropriate interpretation would be that we do not have enough data to make an accurate determination of the effectiveness of interferon. However, all three trials are consistent with a small benefit.

The conceptual advantage of confidence intervals is also seen if we imagine that only the first study, by Mandelli et al.,¹ is available (the first row in Table 1). If you are like me, the fact that this study is of borderline significance ($p = .04$) will bother you. We know that we must draw the line somewhere, but it does not seem right that a small difference in results (say, a few more deaths in the interferon arm) should make a major difference in our recommendations. The confidence interval shows us why: this barely significant study gives us an estimate that is not very precise (a difference between groups of 2% to 38%). If the results had not quite reached statistical significance with $p = .06$, the 95% confidence interval might have been, say, 0% to 39%. This interval must include 0 because the result is not statistically significant; i.e., we cannot be more than 95% sure that there is no difference due to treatment. However, the fact that this interval is *wide* is more important than the fact that it includes 0. In either

case, our estimate *lacks precision*. As our intuition suggests, these results are quite similar in import. How much more useful is it for our patient to know that the difference between the two groups is 2% to 38% rather than 0% to 39%? He is likely to be unimpressed by either estimate.

Think of any intervention that has been subject to a randomized clinical trial and you will see the advantage of confidence intervals over p values. Coronary artery bypass grafting? The p values tell us that it results in better survival than medical treatment for most patients with stable coronary heart disease (5-year survival 10% vs 16%, $p = .0001$; 10-year survival 26% vs 30%, $p = .03$). The confidence intervals tell us that 5-year survival is improved by 3% to 8%, and 10-year survival by 1% to 9%.⁸ Isn't that more informative?

THE PROBLEM WITH SURVIVAL CURVES

Unfortunately, survival data are often analyzed solely by comparing curves and testing for significant differences with p values. This, in fact, is the case for the three interferon trials. Without confidence intervals as survival landmarks (such as the 3-year survival statistics I calculated from the data), we would have to categorize results as positive or negative without being able to evaluate the clinical significance of the findings. In particular, *the meaning of survival differences that are not statistically significant becomes very difficult to determine*. It is extremely important to remember the limitations of so-called negative studies presented this way.

INTERPRETING CLINICAL TRIALS LIKE TEST RESULTS

Another way to look at a clinical trial is to consider it much as one would consider a laboratory test. By this I mean thinking of it as a test that may give false-positive or false-negative results, as well as true results. Let us use as an analogy the use of the antinuclear antibody (ANA) test for diagnosing lupus. Most clinicians know intuitively that the interpretation of an ANA test depends on the setting. A positive ANA test performed as a screening procedure in an asymptomatic patient is likely to be discounted and further testing ordered. Conversely, a positive ANA test in a patient with arthritis, a sun-sensitive rash, and an active urine sediment may be taken at face value as confirmation of the clinical diagnosis of lupus. This intuition is appropriate and is described mathematically by Bayes' theorem, which states that the probability of a disease, given a certain test result, depends not only on the characteristics of the test, but also on the prevalence of the disease in the population.⁹ Thus, the number of positive test results that are incorrect is likely to be greater if the prevalence of the disease in the underlying population is low. This is because there are relatively more patients *without* the disease who have false-positive results than there are patients *with* the disease who have

FIGURE 1. The effect of the prior likelihood of a disease on the ability of a test to predict its presence in a patient. It is assumed that 90% of patients with lupus and 5% of patients without lupus will have a positive ANA test. (A) The ability of a positive ANA test to predict the presence of lupus in a population with a low likelihood of the disease (1% of patients have lupus). (B) The predictive ability of a positive ANA test in a population with a high likelihood of the disease (50% of patients have lupus).

true-positive results. Conversely, a negative test is more likely to be incorrect if the prevalence of the disease is high (see Fig. 1).

Browner and Newman elegantly show how Bayes' theorem can be applied to clinical research.¹⁰ The likelihood that a particular treatment is effective is related to the results of a study and to the probability that the treatment would be effective based on prior information. When a treatment effect is found that seems unlikely, we may disbelieve it: Bayes' theorem tells us that the result is more likely an incorrect one (see Fig. 2). Thus, if we again look only at Mandelli's study in Table 1, whether we recommend interferon treatment to our patient may appropriately depend not only on our interpretation of the confidence interval, but also on our previous beliefs. If we are well versed in the laboratory research on interferons and are convinced that they have an important therapeutic role, we may accept this clinical trial as confirmatory and encourage our patient to proceed with interferon treatment. Conversely, if we have used interferon for years and have seen little more than side effects without clinical benefit, we may rightly discourage our patient and urge restraint until the value of interferon is confirmed. To be fair, of course, we should let our patient know the reasons for our prejudices.

The point here is that, as clinicians, we do not evaluate test results on patients in a vacuum; we consider all data. Why should we treat research results from a clinical

trial any differently? If the result of a clinical trial does not seem to make sense, we need not trust it. Rather, we may wait for confirmation. This is just as reasonable as ordering a repeated ANA test or checking an anti-native DNA antibody on a patient with a positive ANA test. Confirmatory tests are often appropriate; their proper use is one of the signs of a seasoned clinician. Similarly, the appropriateness of confirmatory clinical trials is nicely demonstrated by Parmar et al.¹¹

INTERPRETING MULTIPLE COMPARISONS

A Bayesian approach can also resolve the problem of multiple comparisons, a controversial subject among statisticians. Let me first explain the problem. When many p values (or confidence intervals) are calculated, the probability that one or more significant differences will be found by chance alone increases. As, by definition, there is a 95% chance with each comparison that the p value will be $>.05$ when there is really no difference between the groups compared, the probability of two p values both being $>.05$ is 95% times 95%, or 90%. For three comparisons, all with $p > .05$, the probability is 86%. The probability of *not* all three being $>.05$, i.e., the probability of *at least one* comparison with $p < .05$, is the remainder, 14%. Let us look at how this relates to one of the interferon trials.

Returning again to Mandelli et al.,¹ a closer look reveals that the authors actually performed three separate

A Unlikely hypothesis (10% probability) **B** Likely hypothesis (80% probability)

FIGURE 2. The effect of the prior likelihood of a hypothesis on the ability of clinical trials to predict whether the hypothesis is true or false. It is assumed that the trials are of sufficient size such that 80% of trials of effective therapy will be correctly identified as positive with a $p < .05$. By definition, 5% of trials of ineffective therapy will be incorrectly identified as positive with $p < .05$. P values are used for ease of presentation. The analysis would be the same if we substituted trials with 95% confidence intervals containing only clinically important differences for trials with $p < .05$. (A) The predictive ability of positive trials when unlikely hypotheses are tested (10% of hypotheses are true). (B) The predictive ability of positive trials when likely hypotheses are tested (80% of hypotheses are true).

survival calculations. They looked at survival differences between treatment groups first for all patients, then for only those patients who responded to induction chemotherapy, and finally for those who had only stable disease after induction. Only in the second group, those who responded to induction chemotherapy, was the survival difference significant. However, we have just seen that the likelihood that *at least one* of the three comparisons would have a p value $< .05$, even without a treatment effect, is actually 14%! We have lost our 95% confidence that the survival difference is not due to chance alone. How do we deal with this?

Some statisticians recommend making an adjustment in the significance level in this situation, but this approach increases the likelihood of missing true treatment effects and lacks consistency.¹² If, instead, we use Bayes' theorem to help determine the significance of results, we can more easily discriminate between fishing expeditions and useful findings. In our example, we must decide how plausible it is that survival differences would show up only in the group of patients who responded to chemotherapy. If we accept the authors' implicit belief that this is plausible, the statistically significant result takes on greater importance; i.e., it is more likely a true positive (see Fig. 2). Alternatively, if we think the authors were only hunting for a positive result and there was no

reason for subdividing the patient groups, we may interpret this result very cautiously indeed.

SUMMARY

Let us now return to the patient. Using confidence intervals instead of p values allows us to tell him that interferon treatment may *not* be very important, although the information we have is limited. We can discuss why his friend might have recommended interferon and what factors led us to agree or disagree. We need not make an arbitrary decision. We can use our clinical knowledge, our intuition, and our understanding of our patient, *together* with the results of the clinical trials, to tailor our recommendations to him. In conclusion:

- ◆ Remember that a difference that is not statistically significant is not the same as no difference.
- ◆ Use confidence intervals, not p values, to interpret data, whenever possible. This avoids the problem of false-negative findings and simultaneously provides insight into the precision of results.
- ◆ Allow your clinical knowledge to help put research results into perspective. It is appropriate to interpret clinical trials in conjunction with other information.

The author thanks Dr. Mark Linzer for his encouragement and critical review of the manuscript.

REFERENCES

1. Mandelli F, Avvisati G, Amadori S, et al. Maintenance treatment with recombinant interferon alfa-2b in patients with multiple myeloma responding to conventional induction chemotherapy. *N Engl J Med*. 1990;322:1430-4.
2. Peest D, Deicher H, Coldewey R, et al. A comparison of polychemotherapy and melphalan/prednisone for primary remission induction, and interferon-alpha for maintenance treatment, in multiple myeloma. A prospective trial of the German Myeloma Treatment Group. *Eur J Cancer*. 1995;31A:146-51.
3. Westin J, Rodjer S, Turesson I, et al. Interferon alfa-2b versus no maintenance therapy during the plateau phase in multiple myeloma: a randomized study. *Br J Haematol*. 1995;89:561-8.
4. Ware JH, Delgado F, Donnelly C, Ingelfinger JA. *P values*. In: Bailar JC III, Mosteller F, eds. *Medical Uses of Statistics*. 2nd ed. Boston, Mass: NEJM Books; 1992:181-200.
5. Rothman KJ. A show of confidence. *N Engl J Med*. 1978;299:1362-3.
6. Freiman JA, Chalmers TC, Smith H Jr, Keubler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized clinical trial. Survey of 71 negative trials. *N Engl J Med*. 1978;299:690-4.
7. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Ann Intern Med*. 1994;121:200-6.
8. Yusuf S, Zucker D, Peduzzi P, et al. Effect of coronary artery bypass graft surgery on survival: overview of 10-year results from randomised trials by the Coronary Artery Bypass Graft Surgery Trialists Collaboration. *Lancet*. 1994;344:563-70.
9. Pauker SG, Kassirer JP. Decision analysis. In: Bailar JC III, Mosteller F, eds. *Medical Uses of Statistics*. 2nd ed. Boston, Mass: NEJM Books; 1992:159-79.
10. Browner WS, Newman TB. Are all significant *P* values created equal? The analogy between diagnostic tests and clinical research. *JAMA*. 1987;257:2459-63.
11. Parmar MKB, Ungerleider RS, Simon R. Assessing whether to perform a confirmatory randomized clinical trial. *J Natl Cancer Inst*. 1996;88:1645-51.
12. Rothman KJ. Fundamentals of epidemiologic data analysis: adjustment for multiple comparisons. In: *Modern Epidemiology*. Boston, Mass: Little, Brown and Co.; 1986:147-50.

ANNOUNCEMENT

JGIM Website — Visit us online today!

Please visit the JGIM World-Wide Website:

<http://www.blacksci.co.uk/products/journals/xjgim.htm>