Confidence intervals: from statistical significance to clinical significance

Peter P. Morgan, MD, DPH

The way to a more adequate understanding and treatment of medical data would be opened up if all records, articles and even abstracts gave, besides averages, the numbers of observations and the variation, properly expressed . . ." So wrote Donald Mainland,[1] of Dalhousie University, Halifax, 55 years ago, when introducing the concept of random sampling variation. In a subsequent article Mainland, Du Bilier and Stewart[2] used the example of repeated differential leukocyte counts on the same specimen to show that even after allowing for technical errors sample means varied from batch to batch.

A scan of a run of any medical journal since 1935 will show that clinical researchers were slow to follow Mainland's advice, even if we allow for the fact that many research questions of the earlier era were expressed in descriptive rather than quantitative terms. During the 1940s p values began to appear, and with the efflorescence of computers and the proliferation of databases p values threatened to overwhelm the data. Almost every row and column of data would be compared statistically with its near neighbours or distant relatives. This exercise went beyond hypothesis testing: it was a ruthless search for significance, whose success (almost always based on the magic value of $p < 0.05$) would "validate" the hypothesis, and whose failure would prepare the way for a slightly different study.

As the editors of the new book *Statistics with Confidence*,[3] produced by statistical consultants for the *British Medical Journal*, put it:

> Over the past two or three decades the use of statistics in medical journals has increased tremendously. One unfortunate consequence has been a shift in emphasis away from the basic results towards an undue concentration on hypothesis testing. In this approach data are examined in relation to a statistical "null" hypothesis, and the practice has led to the mistaken belief that studies should aim at obtaining "statistical significance". On the contrary, the purpose of most research investigations in medicine is to determine the magnitude of some factor(s) of interest.

The multiple authors of *Statistics with Confidence* follow in the path of other editorial commentators in the past 10 years who have promoted the use of confidence intervals in displays of statistical analysis. The confidence interval gives a range of values within which the variable of interest is likely to be found, at a specified probability, usually 90%, 95% or 99%. The first chapters of the *BMJ* book are built around the simple illustration of a hypothetical study of systolic blood pressures in men aged 40 to 49 years who have diabetes mellitus and an age-matched group of men who do not. The mean systolic blood pressure of the former is 146.4 mm Hg, of the latter 140.4. The standard deviations are 18.5 and 16.8 mm Hg respectively. If this study had encompassed all men in the age group known to have and not to have diabetes we could confidently assert that those with diabetes have, on average, systolic blood pressures exactly 6 mm higher. But, of course, we are dealing with only a small sample of the universe, so our estimate is uncertain; it is subject to random sampling error.

The uncertainty of the random sampling error is directly related to the inherent variability of the parameter in the population and inversely related to the number in the sample; the latter requirement, since it is more directly under the control of the investigator, is of more concern to us. The larger the sample the less the uncertainty in the estimate of the mean.

In the example, the authors entered the numbers of subjects and the standard deviations for the two populations into a simple formula that gave the standard error of the *difference* between the means, 2.5 mm Hg. The authors then used this value to construct a normal distribution of stan-

Dr. Morgan is consulting editor for CMAJ.

Reprint requests to: Dr. Peter P. Morgan, CMA House, PO Box 8650, Ottawa, Ont. K1G 0G8

dard errors of differences between the means for the 200 men in the sample. In this large population, we know that 1.96 standard errors on either side of the mean encompass 95% of the probabilities under the normal curve, leaving only 2.5% in each tail. This translates to a confidence interval of 9.8, with a lower limit of 1.1 and an upper limit of 10.9. The authors suggest that this information be reported succinctly as follows: "mean = 6.0 mm Hg, 95% confidence interval 1.1 to 10.9, $t = 2.4$, df [degrees of freedom] = 198, P = 0.02".[4]

For the hypothesis tester the essential point in all this is that the critical value of zero, the point of no difference between the blood pressures of the two groups, falls below the lower confidence limit. The *p* test result states the observation in a slightly different way: the probability that the two populations were samples drawn from a single larger population is less than 2 in 100.

A concept aired in the *BMJ* book but perhaps needing emphasis is that both the mean *and* the confidence interval reported in a typical study are only estimates. One should resist the temptation to regard sample means as estimates and confidence intervals as fixed values derived from the population universe. As Snedecor and Cochrane[5] showed, if we draw many repeated samples (with replacement) from the same universe, both the sample means and the upper and lower confidence bounds will be normally distributed. For the 95% confidence intervals in the blood pressure example this signifies that there is a 95% chance that the confidence interval of any given sample, including the one presented, will encompass the true population mean.

It may seem peckish — not to say risky — to disagree with statisticians over nomenclature. However, the now accepted use of the word "interval" is imprecise. The word comes from the Latin *intervallum*, a space between ramparts; its extended meaning is a value or measurement between two points, which it may or may not include. In this editorial, as in the statistical literature, the word is used both correctly to refer to the distance between the two points and incorrectly to refer to the location of the limits themselves. Fear of suggesting that the "limits" include all the data has apparently goaded statistical scriveners into this minor lapse. Why not simply say "95% confidence limits" if that is what is meant? I think the medical world is ready for this.

Although display of confidence intervals does not necessarily interdict display of an accompanying p value, the confidence interval has the following advantages.

- It expresses values not as levels of statistical significance but as actual units — in our example on a scale of millimetres of mercury. The reader can then decide if the lower and upper bounds of 1.1 and 10.9 have clinical significance. Is the estimated mean difference of 6 a clinical concern? Is the clinician worried that there is a 2.5% chance that the true mean difference could be even higher than 10.9? In an article citing this example Langman[6] considered the finding of a difference in blood pressures as great as 10.9 mm Hg "unlikely to be of clinical importance". Bulpitt[7] disagreed. But the point is that confidence intervals allow this sort of debate over statistical results.
- As already noted, the width of the confidence interval allows an estimate of the precision of the sample estimate of the true population mean. Broad confidence intervals can be due to natural variation, but they can always be reduced by increasing the sample size. This is as valuable in planning a trial as it is in interpreting it. As Cox[8] put it, "the standard error should be sufficiently small for us to draw cogent conclusions, but not too small. If the standard error is large the experiment is, by itself, almost useless, whereas an unnecessarily small standard error implies a waste of experimental material."
- The use of the confidence interval is said to de-emphasize hypothesis testing. This depends on the intention of the authors. Confidence intervals can be used to reject a null hypothesis by showing that a difference of zero between means (or a relative risk ratio of 1) lies outside the stipulated bounds. This exercise yields a p value and is the same hypothesis testing that one would see if the p value alone were displayed.
- The use of confidence intervals automatically produces two-tailed tests, favoured by statisticians. However, the argument for using one-tailed tests in certain circumstances has recently been revived.[9] If investigators (and granting agencies) have agreed *a priori* on a one-tailed test of significance it would be sufficient to explain the decision and to present the value of interest in relation to the upper bound.

There are settings in which confidence intervals are redundant. If a statistical analysis is not expected they are obviously superfluous. They are not usually appropriate for simple descriptive data, such as demographic reports or the baseline comparisons of control and study groups in a randomized clinical trial. Tabular presentations of standard errors of the mean, rather than standard deviations, are rarely justified. In graphics standard deviations are again appropriate for descriptive data and confidence intervals for analyses.

Most research published in *CMAJ* deals with means and proportions. However, *Statistics with Confidence* gives instructions on the use of confidence intervals in more complex analyses: those of nonparametric data, relative risks and regression-correlation. The book concludes with statistical guidelines for authors, a useful glossary of the statistical notations in the book, and tables for calculating confidence intervals in *t*, Poisson and various median distributions.

Confidence intervals are not a fundamental departure from the statistical analysis that has been waxing in our journals for the past 50 years. They will not preclude overexamining data or unnecessarily arbitrary hypothesis testing. They are, however, a clear and sensible adjunct to the reader's — and the author's — understanding of research results. They will help to change the focus from statistical significance to clinical significance. And in many instances they will at last satisfy Mainland's ancient plea that variation be properly expressed.

References

1. Mainland D: Chance and the blood count [E]. *Can Med Assoc J* 1934; 30: 656-658
2. Mainland D, Du Bilier B, Stewart CB: The accuracy of differential blood counts. *Can Med Assoc J* 1935; 33: 667-670
3. Gardner MJ, Altman DG (eds): *Statistics with Confidence — Confidence Intervals and Statistical Guidelines, British Medical Journal,* London, 1989: 6
4. Ibid: 14
5. Snedecor GW, Cochrane WG: *Statistical Methods,* 7th ed, Iowa State U Pr, Ames, 1980: 57-59
6. Langman MJS: Towards estimation and confidence limits [E]. *Br Med J* 1986; 292: 716
7. Bulpitt CJ: Confidence intervals. *Lancet* 1987; 1: 494-497
8. Cox DR: *Planning of Experiments,* Wiley, London, 1958: 9
9. Peace KE: The alternative hypothesis: one-sided or two-sided? *J Clin Epidemiol* 1989; 42: 473-476

Bioequivalence of generic aerosol bronchodilators: What are the issues?

Michael Spino, BScPhm, PharmD

On Dec. 9, 1988, the federal government issued a notice of compliance (NOC) to Genpharm Inc., Etobicoke, Ont., for a salbutamol metered-dose inhaler. This is the first time a manufacturer of generic products has received an NOC for a substance administered in this form. The decision raises important issues that require thought and action by the medical and pharmaceutical communities as well as health regulatory agencies.

Bioequivalence

Generic substitution of marketed drugs is now firmly entrenched in the practice of medicine and pharmacy in Canada. For orally administered drugs a pharmacist in most provinces may dispense one brand of a drug, even though the prescription is for another brand, as long as the two drugs meet certain conditions that render them interchangeable. These conditions are usually satisfied if the manufacturer can demonstrate "bioequivalence" of the generic product with the innovator's brand. Bioavailability data are commonly used to substantiate the claim of bioequivalence by manufacturers of generic drugs; this usually means that the rate and extent of absorption of the generic drug and the innovator's product must be comparable (to within 20%). In Canada bioequivalence status has been attained only for certain orally administered drugs and intravenous solutions.

The first step toward interchangeability (which is determined by the provinces) is for the federal government to grant an NOC for the generic product so that the manufacturer can market it anywhere in Canada. The NOC is issued after the Health Protection Branch (HPB), Department of National Health and Welfare, Ottawa, has been satisfied that a manufacturer's submission provides sufficient evidence of safety and efficacy. For most orally administered drugs the HPB has generally accepted a comparative bioavailability study as sufficient evidence. However, since the HPB does not confer the bioequivalence status on products, a company will usually present its submission to provincial regulatory bodies for evaluation of bioequivalence and interchangeability. The company may submit the same bioavailability studies it used to obtain an NOC. If the criteria for bioequivalence are met the status of interchangeability with the innovator's brand would be granted for the product.

NOCs for aerosol bronchodilators

No drugs administered by metered-dose inhal-

Dr. Spino is an associate professor in the faculties of Pharmacy and Medicine, University of Toronto, and is with the Division of Clinical Pharmacology and Toxicology, Hospital for Sick Children, Toronto.

Reprint requests to: Dr. Michael Spino, Faculty of Pharmacy, University of Toronto, 19 Russell St., Toronto, Ont. M5S 2S2