STATISTICAL COMMENT

A brief note on overlapping confidence intervals

Peter C. Austin, PhD,[a,b] and **Janet E. Hux, MD, SM, FRCP(C),**[a,c,d] *Toronto, Ontario, Canada*

Clinical researchers frequently assess the statistical significance of the difference between two means by examining whether the two 95% confidence intervals overlap. The purpose of this brief communication is to illustrate that the 95% confidence intervals for two means can overlap and yet the two means can be statistically significantly different from one another at the $\alpha = 0.05$ level. (J Vasc Surg 2002;36:194-5.)

During seminars in which the results of clinical research are presented, one frequently hears the statement that because the 95% confidence intervals overlap, the means of two different groups are not statistically significantly different from each other (at the $\alpha = 0.05$ level). Furthermore, in the literature, one occasionally observes similar assertions.[1,2] The purpose of this technical note is to discuss the relationship between confidence intervals and hypothesis testing and to illustrate that 95% confidence intervals can overlap, yet the two means can be significantly different from one another at the 0.05 level.

Rosner[3] describes the relationship between hypothesis testing and confidence intervals. In testing of the null hypothesis that a population mean is equal to a specific fixed value (ie, the international normalized ratio is 1.0), the null hypothesis is rejected at a significance level of 0.05 if and only if the 95% confidence interval for the population mean excludes that value. One can make this assertion because the value under the null hypothesis is considered to be fixed. The only source of variability is in the estimation of the population mean with the sample mean.

In testing of the null hypothesis that a mean is equal to a fixed quantity, the only source of variability is in the estimate of the sample mean. Extreme observations are those that lie in the extreme tails of the sampling distribution of the sample mean under the null hypothesis. The probability that a sample mean would lie in the lower 2.5th percentile or the upper 2.5th percentile is 5%. However, when one compares two means, the probability that one mean would lie in the upper 2.5th percentile of that means sampling distribution, while the other simultaneously lies in the lower 2.5th percentile of its sampling distribution, is substantially less than 5%. Hence, despite having overlapping 95% confidence intervals, one can reject the null hypothesis with a *P* value that is substantially less than .05.

In comparison of two groups, the confidence intervals may overlap yet the means may be significantly different from one another. This fact is known in the statistical community[4,5] but bears the occasional repeating within the medical community. Let us assume that we have two independent samples, each composed of n subjects, and that we measure a continuous variable on each subject. For instance, we use 200 patients with diabetes receiving two different drug regimens with hemoglobin A_1c values as the outcome measure. Let $\bar{x}_1$ and $\bar{x}_2$ denote the sample means in the first and second groups, respectively. To simplify the algebra, we assume a common known population variance, σ^2, in each of the two groups. We shall use formulas from Rosner.[3] For simplicity, we assume that the first mean is less than the second mean. Suppose the confidence intervals overlap, with the proportion of the overlap being p. For example, we use mean hemoglobin A_1c of 7.4 (7.0, 7.8) and 8.0 (7.6, 8.4). The width of a 95% confidence interval is equal to $2 \times 1.96\ \sigma/\sqrt{n}$. Then we have that

$$(1)\quad \bar{x}_1 + 1.96\ \sigma/\sqrt{n} = \bar{x}_2 - 1.96\ \sigma/\sqrt{n} + p \times 2 \times 1.96 \times \sigma/\sqrt{n}$$

Rearranging to give the difference between means, we have that

$$(2)\qquad \bar{x}_2 - \bar{x}_1 = 2 \times 1.96 \times \sigma/\sqrt{n} - 2 \times p \times 1.96\,\sigma/\sqrt{n}$$

We can now test the hypothesis that the means are equal in the two groups. We will compute the two-sample z test for independent samples with equal and known variances. The test statistic z is:

$$(3)\qquad z_{test} = (\bar{x}_2 - \bar{x}_1)/\sigma\sqrt{1/n + 1/n}$$

From the Institute for Clinical Evaluative Sciences[a]; the Departments of Public Health Sciences[b] and Medicine,[c] the University of Toronto; and the Division of General Internal Medicine, Clinical Epidemiology Unit and Health Care Research Program, Sunnybrook and Women's College Health Sciences Centre.[d]

Views expressed herein are solely those of the authors and do not represent the views of any of the sponsoring organizations.

Competition of interest: nil.

Reprint requests: Peter Austin, PhD, Institute for Clinical Evaluative Sciences, G-160, 2075 Bayview Ave, Toronto, Ontario, M4N 3M5, Canada (e-mail: peter.austin@ices.on.ca).

0741-5214/2002/$35.00 + 0 **24/1/125015**
doi:10.1067/mva.2002.125015

P values for testing equality of two means when two confidence intervals overlap

Percent overlap of two confidence intervals					
0%	5%	10%	15%	20%	25%
.0056	.0085	.0126	.0185	.0266	.0376

Above table only refers to comparisons of groups with equal sample size and equal variance. Variations would give different results.

We reject the null hypothesis of the equality of the two means if z_{test} is more than 1.96 because the probability that the absolute value of z_{test} is greater than 1.96 is .05. We can now insert the definition of $\bar{x}_2 - \bar{x}_1$ from Eq 2. This results in a test statistic of:

$$z_{test} = \sqrt{2} \times 1.96 \times (1 - p) \quad (4)$$

We will reject the null hypothesis of the equality of the two means when z_{test} is larger than 1.96. This will hold as long as p is less than .29. Hence, as long as the two 95% confidence intervals overlap by less than 29%, one will reject the null hypothesis of the equality of the two means with a *P* value of less than .05. The previous argument can be easily modified to the case in which unknown population variances are estimated with the sample variances. In such a situation, depending on the sample size, the degree of overlap can exceed 29%, and the two means would still be significantly different from one another at the .05 level. The Table contains several degrees of overlap and the *P* values with which one would reject the null hypothesis that the means of the two groups are equal, if the two 95% confidence intervals overlap. Therefore, the fact that two confidence intervals overlap does not necessarily imply that the two means are not significantly different from one another.

We have shown that two 95% confidence intervals can overlap and yet the two means can be statistically significantly different from one another at the $\alpha = 0.05$ level. Hence, one cannot use the fact that two 95% confidence intervals overlap as a substitute for hypothesis testing in assessing the statistical difference between two means. However, one can modify the previous calculations to show that if one constructs 83% confidence intervals, rather than 95% confidence intervals, then if the confidence intervals abut, the *P* value associated with testing the equality of the two means would be approximately .05. Therefore, one can use the criterion of whether or not two 83% confidence intervals overlap as a method for assessing whether or not two means are significantly different from one another at the $\alpha = 0.05$ level.

Returning to the diabetes example, despite the 95% confidence intervals overlapping by 25%, the means differ with $P = .0376$. If the confidence intervals abutted (ie, [7.1, 7.7] and [7.7, 8.3]), the means would differ with $P = .0056$.

In summary, comparing two means is different than comparing one mean with a constant. In comparing two means, there is variability on both measurements of the means, whereas comparing a single mean with a constant involves only one source of variability. Two means may be significantly different from one another, despite the two confidence intervals abutting or having a modest degree of overlap.

REFERENCES

1. Mancuso CA, Peterson MGE, Charlson ME. Comparing discriminative validity between a disease-specific and a general health scale in patients with moderate asthma. J Clin Epidemiol 2001;54;263-74.
2. Sont WN, Zielinski JM, Ashmore JP, Jiang H, Krewski D, Fair ME, et al. First analysis of cancer incidence and occupational radiation exposure based on the National Dose Registry of Canada. Am J Epidemiol 2001;153:309-18.
3. Rosner B. Fundamentals of biostatistics, fourth edition. Belmont, Calif: Duxbury Press; 1995.
4. Goldstein H, Healy MJR. The graphical presentation of a collection of means. J R Stat Soc Assoc 1995;158:175-7.
5. Schenker N, Gentleman JF. On judging the significance of differences by examining the overlap between confidence intervals. Am Statistician 2001;55:182-6.

Submitted Feb 26, 2002; accepted Mar 14, 2002.