Letters
to the Editor

Confidence Intervals vs Significance Tests: Quantitative Interpretation

A paper by Foxman and Frerichs[1] is only the latest example of reports in the Journal in which no tests of significance or p-values for association appear, but instead, confidence intervals for a measure of association are presented. There is no gainsaying that tests of significance have been abused, but at least they have the virtue of providing explicit, pre-specifiable criteria for inferring that an association is real. This is not the case with confidence intervals, at least as far as the paper in question is concerned.

Every single one of the 12 reported 95% confidence intervals (CIs) for the summary odds ratios included the value 1.0, which means that not a single one of the associations was statistically significant at the 0.05 level. Nevertheless, one association was identified as "strong" even though the CI extended from 0.4 to 17.3; a second was identified as "negative" even though the CI extended from 0.08 to 7.8; and a third was identified as "positive" even though the CI extended from 0.3 to 15.4.

I would appreciate learning from the authors, or from the editors or referees who recommended CIs instead of significance tests, just what criteria were employed to conclude that the above (and other) associations were "strong," "negative," and "positive."

REFERENCE
1. Foxman B, Frerichs RR: Epidemiology of urinary tract infection: I. diaphragm use and sexual intercourse. Am J Public Health 1985; 75:1308–1313.

Joseph L. Fleiss, PhD
Professor and Head, Division of Biostatistics, School of Public Health, Columbia University, 600 West 168th Street, New York, NY 10032

Response from Drs. Foxman and Frerichs

We read with interest the letter from Professor Fleiss and welcome the opportunity to respond. In our recent article we report results from a relatively small case-control study [25 cases with primary urinary tract infection (UTI), 25 cases with secondary UTI, and 181 controls]. We presented therein point estimates for two measures of the effect of UTI occurrence among those with the factor of interest to those without the factor: The odds ratio and Mantel-Haenszel summary risk ratio. The reported values (and our descriptive terms, "negative", "strong", "positive" and the like) represent our findings in the single sample we selected from the population of interest. Although we can only speculate as to how bias or unknown confounders might have affected these estimates, we can assess the variability due to the sampling process. Thus we presented 95% confidence intervals for our measures of effect.

Tests of statistical significance reflect both the magnitude of the observed association and the sample size. Given the relatively small sample size of our study, it is not surprising that the 95% confidence intervals we reported were quite wide and, as noted by Fleiss, cover the value 1.0 corresponding to no association. The reader should therefore be cautious in generalizing our results to either the population from which we sampled or other populations of interest. Of course, this same caveat applies to problems of bias or confounding.

Our best guess, however, of the true associations are the point estimates for the odds ratio and Mantel-Haenszel summary risk ratio. Even when not statistically significant, point estimates may significantly add to our understanding. The size and direction of the association can confirm or refute results of previous studies or suggest new hypotheses. In this study, we found a previously unreported association of UTI with diaphragm use (Primary UTI OR = 3.0; 95% CI: 0.9, 11.0 and Secondary UTI OR = 2.3; 95% CI: 0.4, 17.3). Although not statistically significant, the association is consistent with clinical impression and is biologically plausible. Dismissing this association as not real because of lack of statistical significance ignores the other factors commending it. These other factors make the association worthy of note and of the designation as strong.

While we have no aversion to the use of significance tests or p-values (as is historically evident in most of our prior publications), we do feel that relying only on p-values rather than on the size of the parameter estimate to measure the strength of an association has misled many a reader. While the confidence interval may also be misinterpreted, in our case it at least focuses the attention of the reader on the size of the odds or risk ratio rather than on the magnitude of the p-values.

Betsy Foxman, PhD
Assistant Professor, Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI 48109

Ralph R. Frerichs, DVM, DrPH
Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90024

Editors' Note: We have received several other comments akin to Dr. Fleiss' remarks about this paper. We agree with the essence of Dr. Foxman and Dr. Frerichs' response to Dr. Fleiss, and have encouraged the use of confidence intervals. We believe that the quantitative message that they convey is less subject to misinterpretation than significance testing or p-values. The virtue of confidence intervals resides precisely in the fact that they do *not* provide "explicit, pre-specifiable criteria for inferring that an association is real." Such criteria are too rigid and mechanical to be a useful proxy for a quantitative interpretation. Confidence intervals are not a substitute method of significance testing; they convey information about