usual formulation of a causal hypothesis in terms of *counterfactual conditionals* is: "If this sick person had not been exposed, the disease would not have occurred." Applied to a cohort, it is translated: "If these people had not been exposed, the incidence of disease would have been lower." This formulation does not work for case-control studies, because the condition "If the cases had not been exposed" automatically makes the exposure odds zero. But suppose we reformulate the condition as: "If the cases had all been given a preventive that severed the link between exposure and disease." Now the exposure odds among cases (in this counterfactual scenario) would be expected to equal the odds in the denominator population that produced the cases.

In other words, "case-counterfactual" (that is, case-control) studies ask the question: what is the ratio between the observed exposure odds and the expected exposure odds that would have been observed among cases if the effect(s) of exposure had been prevented? Sometimes, an actual control group is not needed to answer this question; for example, among cases of schizophrenia, males substantially outnumber females. We do not need a control group to know that, in the absence of a gender effect (or serious confounding), the sex ratio should be about 1. Likewise, the case-specular design does not need an actual control group if power lines are allocated randomly to this or that side of the street: case-houses' exposure odds (where "exposed" means the power line was on the near side of the street, and "unexposed" means it was on the far side) can be compared with the exposure odds of 1, which is the expected odds if allocation were balanced.

The case-specular design may have only a narrow range of applications, mainly in environmental epidemiology, but it shows that counterfactual definitions of causation have practical, not just theoretical, importance. Careful study of it may inspire the invention of other designs in which controls are counterfactuals.

Malcolm Maclure
Pharmacare, British Columbia Ministry of Health,
1515 Blanshard Street,
Victoria, BC, V8W 3C8 Canada
(address for correspondence) and
Department of Epidemiology,
Harvard School of Public Health, Boston, MA

References

1. Zaffanella LE, Savitz DA, Greenland S, Ebi KL. The residential case-specular method to study wire codes, magnetic fields, and disease. Epidemiology 1998;9:16–20.
2. Greenland S, Robins JM. Identifiability, exchangeability and epidemiologic confounding. Int J Epidemiol 1986;15:413–419.
3. Rothman KJ, Greenland S. Modern Epidemiology. 2nd ed. chapt. 4. Philadelphia: JB Lippincott (in press).
4. Mittleman MA, Maclure M, Sherwood JB, Mulry RP, Tofler GH, Jacobs SC, Friedman R, Benson H, Muller JE. Triggering of acute myocardial infarction onset by episodes of anger. Circulation 1995;92:1720–1725.

That Confounded P-Value

A P-value cannot convey unambiguous information about any relation between exposure and disease. It is inherently confounded information—a mix of information about the size of the effect and the size of the study.[1] Epidemiologists are typically expert in dealing with confounded measures of effect, using standard techniques to factor crude effects explicitly into two components, one due to the effect of the exposure and the other due to the effect of the confounder (or confounders).[2-5] Unfortunately, there has not been similar vigor in disentangling the components of a P-value. It continues to be used mistakenly as a measure of the importance and credibility of study results.

Epidemiology has a longstanding policy of discouraging the use of statistical significance testing, that practice that judges study results according to whether a P-value exceeds or does not exceed a standard yet arbitrary cutoff value.[5-7] Nevertheless, we have not always discouraged the presentation of P-values outside of the context of explicit statistical significance testing. The most common situation for which the reader will encounter P-values in the journal is in the evaluation of trend data. Yet P-values associated with trend data are as confounded as P-values that relate to the difference between two groups.

When editing the article by Cantor and colleagues[8] that appears in this issue, we suggested to the authors that they omit all P-values from the manuscript. The authors agreed to delete most, but they preferred to include P-values for their evaluations of trend. Since we have often published articles that included P-values that were used in much the same way as in the Cantor *et al* study, we felt that, as a matter of fairness and consistency, we should abide by the authors' wishes and allow those P-values in the journal article.

Nevertheless, the discussion prompted us to revisit our editorial policy with regard to reporting P-values. P-values are commonly reported for various tests that relate to epidemiologic analyses, such as a test of the departure of an odds ratio from unity, tests of trend (linear or otherwise), tests of homogeneity, tests of interactions, tests of assumptions underlying the use of

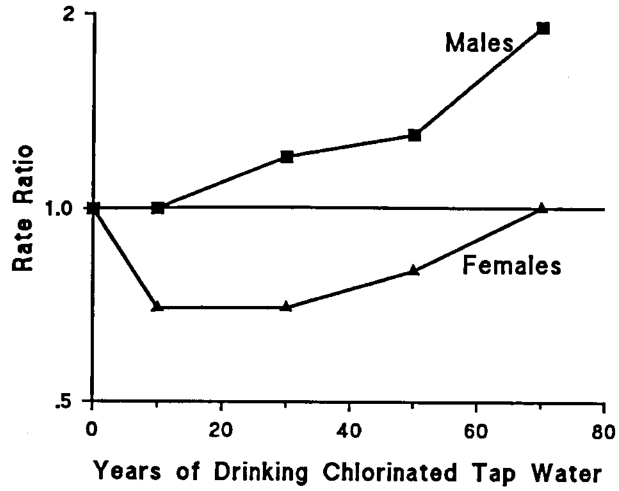

FIGURE 1. Rate ratio of bladder cancer by years of consumption of chlorinated water for males and females.[8]

specific analytical tools, and tests of the comparison of different models, to name a few prominent examples. For all of these, the *P*-value that results from the test is a confounded mix of the magnitude of the underlying measure and the precision of the measure. We believe that we will serve our readers better by discouraging the reporting of confounded information. Therefore, we intend to discourage the reporting of *P*-values in any context in which the confounded elements can be conveniently separated, either numerically, graphically, or otherwise.

For tests of the difference between groups, the practice of summarizing information about effect size and study size with a *P*-value has long been criticized.[5–7] Increasingly in epidemiologic reports, but by no means universally, the two pieces of information that are mixed in the *P*-value are being reported separately. The *size of the effect* is estimated by one or more epidemiologic parameters, such as rate or risk difference, rate or risk ratio, or the proportion of disease attributed to the exposure. The *precision of the estimate*, a function of the size of the study, is described either by a standard error estimate or by reporting a confidence interval around the estimate of effect; the spread of the confidence interval indicates the amount of precision in the estimate.[9–12]

For trend data, one can report an estimate of the slope of a trend line, with its standard error or confidence interval. It is also useful to graph results to examine trends. Scattergrams or smoothed trend lines can depict complicated relations more clearly than *P*-values, which are often based on assumptions the reader cannot easily judge. Consider, for example, the graph presented in Figure 1, showing the trends in risk of bladder cancer for males and females with increasing duration of exposure to chlorinated water from any source, based on data in Table 4 of the article by Cantor *et al.*[8] These graphs convey more information than the *P*-values for trend of 0.002 (males) and 0.88 (females). No one could infer the curves from the *P*-values. Given the curves, no one needs these *P*-values. Rather, it is the shape of the curves, and the precision of their component measurements, that convey the essential information.

Presently, it would be too dogmatic simply to ban the reporting of all *P*-values from **Epidemiology**. However regrettable, the practice of calculating and reporting *P*-values is nearly ubiquitous. In addition, we appreciate that there may be some situations, such as goodness-of-fit evaluations, in which an alternative to the *P*-value is not readily available. Nevertheless, the point remains that the *P*-value is confounded datum, mixing precision with whatever is being measured, be it the fit of a model or the magnitude of a rate ratio. We can tolerate confounded measures when better alternatives are not close at hand, but only reluctantly. By highlighting the confounded nature of *P*-values, we hope to prompt authors to find better ways to separate the core elements of any *P*-value, much the way that point estimates of rate differences or rate ratios and their confidence intervals have already begun to replace *P*-values for the comparison of two rates.

Janet M. Lang
Associate Editor

Kenneth J. Rothman
Editor

Cristina I. Cann
Associate Editor

References

1. Rosenthal R, Rosnow RL. Essentials of Behavioral Research: Methods and Data Analysis. New York: McGraw-Hill, 1984;191.
2. Miettinen OS. Components of the crude risk ratio. Am J Epidemiol 1972; 96:168–172.
3. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic Research: Principles and Quantitative Methods. New York: Life-Time Learning Publications, 1982.
4. Miettinen OS. Theoretical Epidemiology: Principles of Occurrence Research in Medicine. New York: John Wiley and Sons, 1985.
5. Rothman KJ. Modern Epidemiology. Boston: Little, Brown, 1986.
6. Rothman KJ. Significance questing. Ann Intern Med 1986;105:445–447.
7. Rothman KJ. Lessons from John Graunt. Lancet 1996;347:37–39.
8. Cantor KP, Lynch CF, Hildesheim ME, Dosemeci M, Lubin J, Alavanja M, Craun G. Drinking water source and chlorination byproducts. I. Risk of bladder cancer. Epidemiology 1998;1:21–28.
9. Thompson WD. Statistical criteria in the interpretation of epidemiologic data. Am J Public Health 1987;77:191–194.
10. Poole C. Beyond the confidence interval. Am J Public Health 1987;77:195–199.
11. Thompson WD. On the comparison of effects. Am J Public Health 1987; 77:491–492.
12. Poole C. Confidence intervals exclude nothing. Am J Public Health 1987; 77:492–493.