EDITORIAL

Reporting Statistical Information in Medical Journal Articles

STATISTICS IS not merely about distributions or probabilities, although these are part of the discipline. In the broadest sense, statistics is the use of numbers to quantify relationships in data and thereby answer questions. Statistical methods allow the researcher to reduce a spreadsheet of data to counts, means, proportions, rates, risk ratios, rate differences, and other quantities that convey information. We believe that the presentation of numerical information will be enhanced if authors keep in mind that their goal is to clarify and explain. We offer suggestions here for the presentation of statistical information to the readers of general medical journals.

NUMBERS THAT CAN BE OMITTED

Most statistical software packages offer a cornucopia of output. Authors need to be judicious in selecting what should be presented. A chi-square test will typically produce the chi-square statistic, the degrees of freedom in the data, and the *P* value for the test. In general, chi-square statistics, *t* statistics, F statistics, and similar values should be omitted. Degrees of freedom are not needed.

Even the *P* value can usually be omitted. In a study that compares groups, it is customary to present a table that allows the reader to compare the groups with regard to variables such as age, sex, or health status. A case-control study typically compares the cases and controls, whereas a cohort study typically compares those exposed and not exposed. Sometimes authors use *P* values to compare these study groups. We suggest that these *P* values should be omitted. In a case-control or cohort study, there is no hypothesis that the 2 groups are similar. We are interested in a comparison because differences between groups may confound estimates of association. If the study sample is large, small differences that have little confounding influence may be statistically significant. If the study sample is small, large differences that are not statistically significant may be important confounders.[1] The bias due to confounding cannot be judged by statistical significance; we usually judge this based on whether adjustment for the confounding variable changes the estimate of association.[2-6]

Even in a randomized trial, in which it is hoped that the compared groups will be similar as a result of randomization, the use of *P* values for baseline comparisons is not appropriate. If randomization was done properly, then the only reason for any differences must be chance.[7] It is the magnitude of any difference, not its statistical significance, that may bias study results and that may need to be accounted for in the analysis.[8]

Much regression output serves little purpose in medical research publication; this usually includes the intercept coefficient, R^2, log likelihood, standard errors, and *P* values. Estimates of variance explained (such as R^2, correlation coefficients, and standardized regression coefficients (sometimes called effect size) are not useful measures of causal associations or agreement and should not be presented as the main results of an analysis.[9-12] These measures depend not only on the size of any biological effect of an exposure, but also on the distribution of the exposure in the population. Because this distribution can be influenced by choice of study population and study design, it makes little sense to standardize on a measure that can be made smaller or larger by the investigator. Useful measures for causal associations, such as risk ratios and rate differences, are discussed below. Useful measures of agreement, such as kappa statistics, the intraclass correlation coefficient, the concordance coefficient, and other measures, are discussed in many textbooks and articles.[11-20]

Global tests of regression model fit are not helpful in most articles. Investigators can use these tests to check that a model does not have major conflicts with the data, but they should be aware that these tests have low power to detect problems.[21] If the test yields a small *P* value, which suggests a problem with the model, investigators need to consider what this means in the context of their study. But a large *P* value cannot reassure authors or readers that the model presented is correct.[5]

Complex formulas or mathematical notation, such as log likelihood expressions or symbolic expressions for regression models, are not useful for general medical readers.

NUMBERS THAT SHOULD BE INCLUDED

Several authors, including the International Committee of Medical Journal Editors, have urged that research articles present measures of association, such as risk ratios, risk differences, rate ratios, or differences in means, along with an estimate of the precision for these measures, such as a 95% confidence interval.[1,22-29]

Imagine that we compared the outcomes of patients who received treatment A with the outcomes of other patients. If we find that the 2-sided *P* value for this comparison is .02, we conclude that the probability of obtaining the observed difference (or a greater difference) is 1 in 50 if, in the population from which we selected our study subjects, the treatment actually had no effect on the outcomes. But the *P* value does not tell read-

Data From a Hypothetical Randomized Controlled Trial of Treatment A Compared With Standard Care

	No. (%) of Patients		
Outcome	Treatment A (N = 800)	Standard Care (N = 1200)	Risk Ratio (95% Confidence Interval)
Death	18 (2.3)	54 (4.5)	0.50 (0.30 to 0.85)
Hospitalization	100 (12.5)	230 (19.2)	0.65 (0.52 to 0.81)
Treatment complications*			
None	563 (75.1)	615 (55.0)	Reference
1	113 (15.1)	280 (25.0)	0.53 (0.44 to 0.65)
2 or more	74 (9.9)	223 (19.9)	0.44 (0.34 to 0.56)

*Data on complications were missing for 6.3% of the treatment A patients and 6.8% of the standard care patients.

ers if those who received treatment A did better or worse compared with those who did not. Nor does it tell readers how much better or worse one group did compared with the other. However, if we report that the risk ratio for a bad outcome was 0.5 among those who received treatment A, compared with others, readers can see both the direction (beneficial) and the size (50% reduction in the risk of a bad outcome) of treatment A's association with bad outcomes. It is also useful, when possible, to show the proportion of each group that had a bad outcome or to use something similar, such as a Kaplan-Meier survival curve. If we report that the 95% confidence interval around the risk ratio of 0.5 was 0.3 to 0.7, readers can see that the null hypothesis of no association (a risk ratio of 1.0) is unlikely and that risk ratios of 0.9 or 0.1 are also unlikely. If we report that the risk ratio was 0.5 (95% confidence interval, 0.2 to 1.3), a reader can see that the estimate of 0.5 is imprecise and the data are compatible with no association between treatment and outcome (a risk ratio of 1.0) and are even compatible with a harmful association (a risk ratio greater than 1.0). A point estimate and confidence interval convey more information than the *P* value for a test of the hypothesis of no association. Similarly, means can be compared by presenting their differences with a 95% confidence interval for the difference.

We acknowledge that sometimes *P* values may serve a useful purpose,[30] but we recommend that point estimates and confidence intervals be used in preference to *P* values in most instances. If *P* values are given, please use 2 digits of precision (eg, *P*=.82). Give 3 digits for values between .01 and .001 and report smaller values as *P* <.001. Do not reduce *P* values to "not significant" or "NS."

DESCRIPTIVE TABLES

In tables that compare study groups, it is usually helpful to include both counts (of patients or events) and column percentages (**Table**). In a case-control study, there are usually column headings for the cases and controls. For clinical trials or cohort studies, the column headings are typically the trial arms or the exposure categories. Listing column percentages allows the reader to easily compare the distribution of data between groups. Do not give row percentages.

In tables of column percentages, do not include a row for counts and percentages of missing data. Doing this will distort the other percentages in the same column, making it difficult for readers to compare known information in different columns. The records with missing data are best omitted for each variable. The investigator hopes that the distribution of information about those with missing data was similar to those with known data. The amount of missing data should be described in the methods section. If there is a lot of missing data for a variable, say more than 5%, a table footnote can point this out (Table).

ODDS RATIOS VS RISK RATIOS

In a case-control study, associations are commonly estimated using odds ratios. Because case-control studies are typically done when the study outcome is uncommon in the population from which the cases and controls arose, odds ratios will approximate risk ratios.[31,32] Logistic regression is typically used to adjust odds ratios to control for potential confounding by other variables.

In clinical trials or cohort studies, however, the outcome may be common. If more than 10% of the study subjects have the outcome, or if the baseline hazard of disease is common in a subgroup that contributes a substantial portion of subjects with the outcome, then the odds ratio may be considerably further from 1.0 than the risk ratio.[6,33] This may result in a misinterpretation of the study results by authors, editors, or readers.[34-40] One option is to do the analysis using logistic regression and convert the odds ratios to risk ratios.[41-43] Another option is to estimate a risk ratio using Poisson regression, negative binomial regression, or a generalized linear model with a log link and binomial error distribution.[44-55] Whatever choice is made, we urge authors not to interpret odds ratios as if they were risk ratios in studies where this interpretation is not warranted.

POWER CALCULATIONS AFTER THE RESULTS ARE KNOWN

Reporting of power calculations makes little sense once the study has been done.[56,57] We think that reviewers who request such calculations are misguided. We can never know what a study would have found if it had been larger. If a study reported an association with a 95% confidence interval that excludes 1.0, then the study was not underpowered to reject the null hypothesis using a 2-sided significance level of .05. If the study reported an association with a 95% confidence interval that includes 1.0, then by that standard the data are compatible with the range of associations that fall within the confidence interval, including the possibility of no association. Point estimates and confidence intervals tell us more than any power calculation about the range of results that are compatible with the data.[58] In a review of this topic, Goodman and Berlin wrote that " . . . we cannot cross back over the divide and use pre-experiment numbers to interpret the result. That would be like trying to convince someone that buying a lottery ticket was foolish (the before-experiment perspective) after they hit the lottery jackpot (the after-experiment perspective)"[59(p201)]

CITATIONS FOR METHODS SECTIONS

In the methods section, authors should provide sufficient information so that a knowledgeable reader can understand how the quantitative information in the results section was generated. For common methods, such as the chi-square test, Fisher exact test, the 2-sample *t* test, linear regression, and logistic regression, we see no need for a citation. For proportional hazard models, Poisson regression, and other less common methods, we recommend that a textbook be cited so that an interested person could read further.

Authors sometimes state that their analytic method was a specific command in a software package. This is not helpful to persons without that software. Tell readers the method using statistical nomenclature and give appropriate citations to statistical textbooks and articles, so that they could implement the analysis in the software of their choice.

We see no reason to mention or cite the software used for common statistical methods. For uncommon methods, citing the software may be helpful because the reader may want to acquire software that implements the described method and, for some newer methods, results may be somewhat different depending on the software used. If you are in doubt, we suggest citing your software.

CLARITY VS STATISTICAL TERMS

Because the readers of general medical journals are not usually statisticians, we urge that technical statistical terms be replaced with simpler terms whenever possible. Words such as "stochastic" or "Hessian matrix" are rarely appropriate in an article and are never appropriate in the results section.

As an example, imagine that we have done a randomized trial to estimate the risk ratio for pneumonia among those who received a vaccine compared with others. Study subjects ranged in age from 40 to 79 years. We used regression to estimate that the risk ratio for pneumonia was 0.5 among vaccine recipients compared with controls. As part of our analysis, we wanted to know if this association was different among those who were younger (40 to 59 years) compared with those who were older (60 to 79 years). To do this, we introduced what statisticians call an interaction term between treatment group and age group. It is fine to say this in the methods. But in the results we can avoid both the statisticians' term (interaction) and the epidemiologists' term (effect modification) and simply say, "There was no evidence that the association between being in the vaccine group and pneumonia varied by age group (P=.62)."

Accurate statements about statistical methods will sometimes require words that will be unfamiliar to some readers. We are not asking for clarity at the expense of accuracy, and we appreciate that sometimes part of the methods section will be beyond the general reader. The results section, however, must be written so that the average reader can understand the study findings.

COMMON LANGUAGE PITFALLS

Avoid use of the word "significant" unless you mean "statistically significant"; in that case, it is best to use both those words.

Do not confuse lack of a statistically significant difference with no difference.[60] Imagine that the mean age is 38.3 years in group A and 37.9 years in group B, with a mean difference of 0.4 years (95% confidence interval, 2.4 to −1.6). Do not say that the 2 groups did not differ in regard to age; they clearly do differ, with a mean difference of 0.4 years. It might be reasonable to say that the 2 groups were similar with regard to age or that differences in mean age were not statistically significant.

DOGMA VS FLEXIBILITY

Biostatistics, like the rest of medicine, is a changing field. Nothing we have said here is fixed in stone. Today, for example, we recommend confidence intervals as estimates of precision, but we would be quite willing to accept a manuscript with likelihood intervals instead.[5,61-63] If authors think they have a good reason to ignore some of our recommendations, we encourage them to write their manuscript as they see fit and be prepared to persuade and educate reviewers and editors. If authors keep in mind the goals of clarity and accuracy, readers will be well served.

Peter Cummings, MD, MPH
1524 Bear Creek Dr
Bishop, CA 93514
(e-mail: peterc@u.washington.edu)
Frederick P. Rivara, MD, MPH
Seattle, Wash

REFERENCES

1. Lang JM, Rothman KJ, Cann CI. That confounded *P* value. *Epidemiology.* 1998; 9:7-8.
2. Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *Am J Epidemiol.* 1989;129:125-137.
3. Maldonado G, Greenland S. Simulation study of confounder selection strategies. *Am J Epidemiol.* 1993;138:923-936.
4. Kleinbaum DG. *Logistic Regression: A Self-Learning Text.* New York, NY: Springer-Verlag; 1992:168.
5. Rothman KJ, Greenland S. *Modern Epidemiology.* Philadelphia, Pa: Lippincott-Raven; 1998:195-199, 255-259, 410-411.
6. Greenland S, Morgenstern H. Confounding in health research. *Annu Rev Public Health.* 2001;22:189-212.
7. Matthews JNS. *An Introduction to Randomized Controlled Clinical Trials.* London, England: Arnold; 2000:64-65.
8. Rothman KJ. Statistics in nonrandomized studies. *Epidemiology.* 1990;1:417-418.
9. Greenland S, Schlesselman JJ, Criqui MH. The fallacy of employing standardized regression coefficients and correlations as measures of effect. *Am J Epidemiol.* 1986;123:203-208.
10. Greenland S, Maclure M, Schlesselman JJ, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. *Epidemiology.* 1991;2:387-392.
11. Altman DG. *Practical Statistics for Medical Research.* New York, NY: Chapman & Hall; 1991:396-419.
12. van Belle G. *Statistical Rules of Thumb.* New York, NY: John Wiley & Sons; 2002: 56-68.
13. Fleiss JL. *Statistical Methods for Rates and Proportions.* New York, NY: John Wiley & Sons; 1981:212-236.

14. Nelson JC, Pepe MS. Statistical description of interrater variability in ordinal ratings. *Stat Methods Med Res*. 2000;9:475-496.
15. Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med*. 2002;21:2109-2129.
16. Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*. Oxford, England: Blackwell Science; 2002:704-707.
17. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255-268.
18. Lin LI. A note on the concordance correlation coefficient. *Biometrics*. 2000;56: 324-325.
19. Bland JM, Altman DG. Statistical methods for assessing agreement between 2 methods of clinical measurement. *Lancet*. 1986;1:307-310.
20. Bland JM, Altman DG. Measurement error. *BMJ*. 1996;313:744.
21. Hosmer DW, Hjort NL. Goodness-of-fit processes for logistic regression: simulation results. *Stat Med*. 2002;21:2723-2738.
22. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *BMJ (Clin Res Ed)*. 1986;292:746-750.
23. Rothman KJ. Significance questing. *Ann Intern Med*. 1986;105:445-447.
24. Walker AM. Reporting the results of epidemiologic studies. *Am J Public Health*. 1986;76:556-558.
25. Savitz DA. Is statistical significance testing useful in interpreting data? *Reprod Toxicol*. 1993;7:95-100.
26. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *N Engl J Med*. 1991;324:424-428.
27. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *JAMA*. 1997;277:927-934.
28. The value of *P*. *Epidemiology*. 2001;12:286.
29. Poole C. Low *P*-values or narrow confidence intervals: which are more durable? *Epidemiology*. 2001;12:291-294.
30. Weinberg CR. It's time to rehabilitate the *P*-value. *Epidemiology*. 2001;12:288-290.
31. Kelsey JL, Whittemore AS, Evans AS, Thompson WD. *Methods in Observational Epidemiology*. New York, NY: Oxford University Press; 1996:36.
32. MacMahon B, Trichopoulos D. *Epidemiology: Principles and Methods*. Boston, Mass: Little, Brown & Co; 1996:169-170.
33. Greenland S. Interpretation and choice of effect measures in epidemiologic analyses. *Am J Epidemiol*. 1987;125:761-768.
34. Relman AS. Medical insurance and health: what about managed care? *N Engl J Med*. 1994;331:471-472.
35. Welch HG, Koepsell TD. Insurance and the risk of ruptured appendix [letter]. *N Engl J Med*. 1995;332:396-397.
36. Altman DG, Deeks JJ, Sackett DL. Odds ratios should be avoided when events are common [letter]. *BMJ*. 1998;317:1318.
37. Altman D, Deeks J, Sackett D. Odds ratios revisited [letter]. *Evid Based Med*. 1998; 3:71-72.
38. Sackett DL, Deeks JJ, Altman DG. Down with odds ratios! *Evid Based Med*. 1996; 1:164-166.
39. Schwartz LM, Woloshin S, Welch HG. Misunderstanding about the effects of race and sex on physicians' referrals for cardiac catheterization. *N Engl J Med*. 1999; 341:279-283.
40. Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med*. 2002;21:1575-1600.
41. Zhang J, Yu KF. What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *JAMA*. 1998;280:1690-1691.
42. McNutt LA, Hafner JP, Xue X. Correcting the odds ratio in cohort studies of common outcomes [letter]. *JAMA*. 1999;282:529.
43. Nelder JA. Statistics in medical journals: some recent trends [letter]. *Stat Med*. 2001;20:2205.
44. Breslow NE, Day NE. *The Design and Analysis of Cohort Studies*. Lyon, France: International Agency for Research on Cancer; 1987:120-176. *Statistical Methods in Cancer Research*; vol 2.
45. McCullagh P, Nelder JA. *Generalized Linear Models*. New York, NY: Chapman & Hall; 1989.
46. Gourieroux C, Monfort A, Tognon C. Pseudo-maximum likelihood methods: theory. *Econometrica*. 1984;52:681-700.
47. Gourieroux C, Monfort A, Tognon C. Pseudo-maximum likelihood methods: applications to Poisson models. *Econometrica*. 1984;52:701-720.
48. Wacholder S. Binomial regression in GLIM: estimating risk ratios and risk differences. *Am J Epidemiol*. 1986;123:174-184.
49. Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychol Bull*. 1995; 118:392-404.
50. Long JS. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, Calif: SAGE Publications; 1997:217-250.
51. Cameron AC, Trivedi PK. *Regression Analysis of Count Data*. New York, NY: Cambridge University Press; 1998.
52. Lloyd CJ. *Statistical Analysis of Categorical Data*. New York, NY: John Wiley & Sons; 1999:84-87.
53. Cummings P, Norton R, Koepsell TD. Rates, rate denominators, and rate comparisons. In: Rivara FP, Cummings P, Koepsell TD, Grossman DC, Maier RV, eds. *Injury Control: A Guide to Research and Program Evaluation*. New York, NY: Cambridge University Press; 2001:64-74.
54. Hardin J, Hilbe J. *Generalized Linear Models and Extensions*. College Station, Tex: Stata Press, 2001.
55. Wooldridge JM. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, Mass: MIT Press; 2002:646-649.
56. Bacchetti P. Author's thoughts on power calculations [letter]. *BMJ*. 2002;325: 491.
57. Senn SJ. Power is indeed irrelevant in interpreting completed studies [letter]. *BMJ*. 2002;325:1304.
58. Smith AH, Bates MN. Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology*. 1992;3:449-452.
59. 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;121:200-206.
60. Altman DG, Bland MJ. Absence of evidence is not evidence of absence. *BMJ*. 1995;311:485.
61. Royall R. *Statistical Evidence: A Likelihood Paradigm*. Boca Raton, Fla: CRC Press; 1997.
62. Goodman SN. Toward evidence-based medical statistics, 1: the *P* value fallacy. *Ann Intern Med*. 1999;130:995-1004.
63. Goodman SN. Toward evidence-based medical statistics, 2: the Bayes factor. *Ann Intern Med*. 1999;130:1005-1013.