

Current reviews of allergy and clinical immunology

(Supported by a grant from GlaxoSmithKline, Inc, Research Triangle Park, NC)
Series editor: Harold S. Nelson, MD

Statistical errors in immunologic research

James R. Murphy, PhD *Denver, Colo*

This activity is available for CME credit. See page 22A for important information.

Medical research articles have always been subject to errors in reporting statistical results. Although most of these are minor, they raise questions about the integrity of medical research. Most of the errors come from a misunderstanding about the tools used in statistical analysis. This article discusses some of the most frequent errors and provides examples of how to deal with them correctly. (*J Allergy Clin Immunol* 2004;114:1259-63.)

Key words: Statistics, *P* value, power, errors in reporting, variance

Recently, Olsen¹ published a commentary indicating the number and type of statistical errors that she found in reviewing articles in *Infection and Immunology*. She also referenced several articles indicating similar problems in medical journals in other specialty areas. In a related paper in *BMC Medical Research Methodology*, Garcia-Berthou and Alcaraz² noted that almost 12% of the articles reviewed in *Nature* and the *British Medical Journal* contained *P* values that did not match the values of the test statistics given. Given that these same problems have been previously reported at least as far back as 1979³ and appear to be ubiquitous, this article presents a more in-depth discussion of the problems, paying particular attention to how they might affect research in the areas of allergy and immunology.

All of the problems hinge on the understanding of what a statistical test is doing and what a *P* value means. Therefore this article starts with a brief description. There are a variety of statistical tests used for comparison, modeling, prediction, and reduction of data. The errors discussed in the Olsen¹ article deal with problems when statistics are used to compare groups, and this article will deal exclusively with that area.

BASIC DEFINITIONS

A statistical test is a procedure to determine whether a defined quantity is larger (smaller) than you should expect by chance. The major effect of the test comes from the way in which the study is designed (ie, set up and executed) to get you to a point where you can do the test. Although readers might focus a great deal on whether a test result is significant, this is really not very important unless the study is designed properly in the first place. A statistical test is therefore somewhat different from a laboratory test in which a change of color or light intensity might be all that is needed to detect a physiologic change, although even here you need to have proper design and procedure to make sure that contaminants have not produced your result.

The indicator that some change has taken place greater than the level of chance is the *P* value. There is consensus in medical research that a *P* value of less than .05 means that the statistical test you are conducting is significant (ie, that the result you see is not just the result of chance). It is worth noting again that a significant *P* value does not indicate a physiologic or chemical change in a process. Instead it indicates that the accumulation of data by consensus has reached a level that supports whatever change you proposed in your design. The *P* level you achieve in a study is only one value on a continuous scale of *P* values. In immunologic research its closest counterpart is probably a light intensity reading taken from a microarray in which the .05 level of significance is similar to the 2-fold difference in intensity required to indicate that upregulation or downregulation has occurred.

Focusing on a significant *P* value provides a certain economy for making decisions when you are examining a large number of situations in which decisions need to be made, such as examining a number of articles for possible inclusion in a journal.

However, the difference between a *P* value of .051 (nonsignificant) and a *P* value of .049 (significant) might mean nothing at all in terms of the underlying physiology or chemistry. Significant *P* values need to be supported with additional logic and secondary evidence to be truly meaningful. It might also be useful to think of the *P* value as an indication of how likely you are to be able to replicate your current result rather than as a signal that

From the Division of Biostatistics, National Jewish Medical and Research Center.

Disclosure of potential conflict of interest: J. R. Murphy—none disclosed.

Received for publication September 9, 2004; revised September 17, 2004; accepted for publication September 21, 2004.

Reprint requests: James R. Murphy, PhD, Division of Biostatistics, National Jewish Medical and Research Center, 1400 Jackson St, Denver, CO 80206.

E-mail: murphyj@njc.org.

0091-6749/\$30.00

© 2004 American Academy of Allergy, Asthma and Immunology

doi:10.1016/j.jaci.2004.09.023

change has occurred. The smaller the P value, the less likely the result is due to chance, and the more likely you will be able to replicate it.

The power of a study, which is a property of its sample size, indicates how likely you are to be able to detect the size of difference (change) you have specified in the design. Taken together, the statistical test, P value, and power can also be understood in the context of the false-positive and false-negative rates of a screening or diagnostic test. The statistical test is the test, the .05 level is the false-positive rate you are willing to accept, and the power is the true-positive rate of your test if the change you are specifying is correct. Adopting this view might also help to explain that a nonsignificant result on a statistical test does not mean that you have disproved the null hypothesis. It is more likely to mean that because of inadequate sample size or imprecision in your measurements, you are getting more false-negative results than you expected.

It is important to note that if I would like to show that the difference between 2 groups is not significant, all I have to do is to design a study with too small a sample size and that result is guaranteed. Conversely, if I want to show that something is significant, all I have to do is get a large enough sample size and I can prove that. The logic and science of the process is in selecting a value for change that is meaningful in a scientific or clinical sense and then specifying a statistical design and testing structure that allows you to demonstrate this meaningful difference with the smallest number of subjects. The design and testing structure of the study, including the major statistical results, should be found in the abstract at the beginning of the article.

Below I list Olsen's main problem areas with brief discussions. After each problem, I give examples of acceptable ways to deal with the problem in a succinct manner. These examples are taken from articles by Strunk et al,⁴ Asero et al,⁵ Adams et al,⁶ Sorensen et al,⁷ and Eggesbe et al⁸ in recent issues of the JACI. The quotes used in the current article are usually taken from a section of the quoted articles labeled "Statistical methods" or "Statistical analysis," although Sorensen et al⁷ use the term "Data analysis." All of these articles avoid the errors listed below and could be used as good examples of the correct way to report statistics in the research literature.

I have used the abstract of Strunk et al⁴ as a good example of what you should find at the beginning of an article. His abstract gives a good, concise description of the objectives and design of his study. All of the articles cited have similar descriptions that immediately allow the reader to determine what they should be looking for in the article itself. In the case of Strunk et al,⁴ the reader knows that in the article they should be seeing a discussion of correlation, rank regression, randomization, and a discussion of definitions for mild and moderate asthma.

Objective: The aim of this study was to find correlations between eNO and other characteristics of children with mild to moderate asthma currently not taking medications.

Methods: Children aged 6 to 17 years with mild to moderate persistent asthma, taking only albuterol as needed, were characterized during 2 visits 1 week apart before being randomly assigned into a clinical trial. At the screening visit, online measurements of eNO, spirometry before and after bronchodilator, and biomarkers of peripheral blood eosinophils, serum eosinophil cationic protein, total serum IgE, and urinary leukotriene E4 were obtained. During a week characterization period before randomization, symptoms were recorded on a diary and peak expiratory flows were measured twice daily using an electronic device. At the randomization visit, eNO was repeated followed by a methacholine challenge and aeroallergen skin testing. Correlations and rank regression analyses between eNO and clinical characteristics, pulmonary function and biomarkers were evaluated.

EXAMINING THE ERRORS

It is quite true that some of the errors Olsen¹ found, items 1 to 6 below, could be the result of editorial policies that restricted some of the detail given in the original articles. Because removing these errors is critical to the correct interpretation of statistical results, I hope that editorial problems are becoming less of an issue. Recent guidelines for publications in biomedical journals would suggest that the type of detail suggested in Olsen's article should be accepted in the future.⁹

1. *Failure to document the statistical method used or using an incorrect method.* Like any test, statistical tests work best in the specific situations for which they are constructed. Because most statistical tests use a P value of .05 or less as their significance indicator, it might seem like they are interchangeable, but differences in the type of data used, the variability of the data, and the study design dictate that there is an appropriate test that should be used for each situation. Statistical software packages have made it easy to try a variety of tests in situations in which they might or might not apply. Using the incorrect test is similar to performing a laboratory test without using proper controls. You can get a significant result, but you will not know what it means. If you are uncertain of what type of test to use, it is best to consult a statistician.

In his "Statistical methods" section, Strunk et al⁴ list 2 tests that were used for the analyses in the article. These are the Spearman correlation coefficient as a measurement of the main objective and stepwise rank regression for further modeling. The results, tables, and figures are all clearly related to one of the procedures mentioned.

Asero et al⁵ note the test used, a correction for it, and the probability level considered significant. "Proportions were compared by using the chi-square test with the Yates correction. Probability level of less than 5% were considered statistically significant."

Adams et al⁶ note that a P value of .05 or less is not the only criteria that they use for significance. "Multiple logistic regression was used to assess independent effects in models for anti-inflammatory therapy use. Variables significant at the 0.10 level in bivariate analysis were

entered simultaneously into a multiple logistic regression model.”

Eggesbe et al⁸ make a comment about the power of their study for certain outcomes. As mentioned above, making a note of the power or lack thereof is as important as determining the significance of the test results.

The small groups make the study prone to type II errors (ie, not detecting factors that might be of importance). This uncertainty affects, among other results, the results concerning antibiotics. Furthermore, it might be the reason why the association between egg allergy and cesarean delivery failed to reach significance in the strata of allergic mothers in spite of the strength of the observed association.

To determine the correctness of these tests, you need to determine that they are appropriate for the objective of the study, the design used, and the type of data being used. Part of the job of the statistical reviewer of a journal is to make sure that the statistical procedures published in that journal meet the above criteria.

2. *Failure to adjust for multiple comparisons (multiple tests of the same data).* In the same way that clinical tests can produce false-positive and false-negative results, statistical tests can do the same. As in clinical testing, the more statistical tests you do, the more likely you are to get a positive (true or false) result at some point. In the clinical setting the potentially misleading results can be discounted if they do not fit the developing pattern of the differential diagnoses. In statistical testing an adjustment is made to the *P* value's level of significance so that the more tests you do, the more difficult it is to get a significant result. Even among statisticians, there is disagreement about when this adjustment needs to be made and exactly how much of an adjustment needs to be made. There is agreement, however, that results that are obtained as a result of frequent examination of the data are qualitatively different from results that come about because of a prestated hypothesis. At a minimum, an author should inform his audience when this additional data examination has been necessary to provide a significant result. The audience is then free to question whether it is likely that this result could be replicated in a similar study.

Sorensen et al⁷ report their multiple comparison adjustment as follows.

Post hoc multiple comparisons were only preformed on variables and challenges for which significant differences between the CFS and control groups were found, including group-by-time interaction. The Tukey-Kramer multiple comparison method was used.

There are several types of adjustments in addition to the Tukey-Kramer that can be used for multiple comparisons. For a very practical discussion of this issue, see Curran-Everett.¹⁰

3. *Reporting observations without a statistical test.* Significance, like beauty, might be in the eye of the beholder. Although it is true that there are situations in which a unique observation is all that is needed to make your case, these are becoming increasingly rare. Graphic representations can be misleading, and large differences between groups that come with large variability might not

be significant, no matter how they look. To really make the case that your result is larger than chance requires the appropriate statistical test.

Strunk et al⁴ present graphic scatter plots in their Fig 2 that suggest a relationship between exhaled nitric oxide and certain other variables. However, they make the case for these relationships not on the basis of the graphic but on the basis of the statistical test for correlation, which they presents in the “Biomarkers results” section.

Sorensen et al⁷ present plots with standard error bars in Fig 1, A and B, which do seem to indicate differences. They make their point, however, by referring to the significance of the test results in the “Study data” section.

4. *Using a statistical test that requires an underlying normal distribution on data that are not normally distributed.* The normal (also known as the Gaussian or bell curve) distribution has some very nice properties that make it easy to combine measurements and calculate test results and *P* values. Before the advent of computers, these properties made it worthwhile to try to transform data so that it was normally distributed and to concentrate on tests that were created for a normal distribution. It is also the case that a normal distribution might be a good approximation to how a large group of observations are distributed. Knowing the distribution (particularly knowing the mean and variance and their relation to one another) of your measurements puts you into a category of tests called parametric. There are other distributions in this category in addition to the normal, but the normal is the one for which most statistical tests were originally created.

Unfortunately, much biomedical research data are not normally distributed, and using tests designed for normally distributed data on this type of data can cause misleading results. Fortunately, most tests requiring normality have nonparametric counterparts that make very minimal assumptions about how the data are distributed. Most statistical packages contain a series of nonparametric tests that can be used in place of or along with parametric tests. These tests usually have slightly lower powers than their normal counterparts, but they give more correct results.

An alternative to using a nonparametric test is to transform the data so that it becomes more normal. A popular transformation is to take the natural logarithm of your measurements and create what is called a log normal distribution. Here again there are a number of transformations that can be used, and it is important to test the results to make sure that the resultant observations are normal. You cannot simply assume that the transformation has worked the way you expected.

Strunk et al⁴ identify nonnormal distributions and use both nonparametric tests and transformations where they are appropriate. The following is from his “Statistical methods” section. Note that the Spearman test is a nonparametric test for correlation.

Spearman correlation coefficients were used to assess the relationships among the biomarkers, clinical and pulmonary function measures, and allergy skin test reactivity. In this

TABLE II. Characteristics of the 144 CLIC participants: Demographics, pulmonary function test and clinical characteristics and skin test reactivity

Female	41%		
Minority	48%		
	Mean $\pm$ SD	Median	Lower and upper quartiles
Age y	114 $\pm$ 3.4	11.42	8.0, 14.3
Symptoms during characterization, days per wk	4.1 $\pm$ 2.5	4.4	2.0, 6.2
Spirometry: forced vital capacity, % predicted	105.3 $\pm$ 11.8	104.8	97.4, 113.0
FEV ₁ * % predicted	95.0 $\pm$ 12.9	95.3	85.8, 103.1
FEV ₁ /forced vital capacity, %	79.4 $\pm$ 8.5	81.0	72.0, 86.0
Bronchodilator reversibility, % change in FEV ₁ from baseline	15.3 $\pm$ 9.4	14.4	10.0, 19.9
PC ₂₀ † mg/ml	1.2, 3.8†	1.1†	0.45†, 3.4†
AM PEF, % predicted	77.2 $\pm$ 12.7	77.5	68.4, 85.4
PEF variability, %‡	9.5 $\pm$ 5.8	8.4	5.6, 11.9
Skin test reactivity, No. positive tests of 8 tested	2.8 $\pm$ 2.3	3.0	1.0, 4.0

(Reproduced with permission from: Strunk RC, Szefer SJ, Phillips BR, Zeiger RS, Chinchilli VM, Larsen G, et al. Relationship of exhaled nitric oxide to clinical and inflammatory markers of persistent asthma in children. *J Allergy Clin Immunol* 2003;112(5):883-892.)

*Geometric mean, coefficient of variation

†Median and 1st and 3rd quartiles are reported on the log₂ scale

‡AM PEF-PM PEF/[(AM PEF + PM PEF)/2] $\times$ 100

manner, potential predictors of eNO were identified. PC₂₀ was analyzed on the log base 2 scale because of doubling doses in the methacholine challenge procedure. The eNO measurements and many of the biomarkers displayed a skewed distribution, and these were analyzed on the logarithmic scale.

Sorensen et al⁷ note the nonnormality of their data and the need for transformation in the "Data analysis" section. "The data were generally highly right-skewed, thus natural log-transformed variables were analyzed."

5. *Not identifying or properly labeling the type of variance estimate being used.* There are a variety of variance structures that might apply, but the 2 that are most commonly not labeled or mislabeled are the SD and the SE. The SD measures the variability in your original data. The SE measures the variability in the mean or average value of your data. When you want to demonstrate how variable your mean is (this is important for reproducibility and statistical testing), you want to use the SE and corresponding SE bars around the mean. When you want to show how variable a single observation could be expected to be, you want to use the SD and error bars based on the SD around a single measurement. You also need to state whether you are using 1, 2, or sometimes 3 times the SE or SD. Each one of these limits encompasses a different proportion of the variables you are examining. For example, a 95% CI around a mean taken from normally distributed data is obtained by taking $\pm 2 \times \text{SE}$. Using 1 instead of 2 gives you a 68% CI.

In their Table II, Strunk et al⁴ clearly indicate that they are providing the mean $\pm$ SD. What they want to demonstrate is the variability of his single observations. In this presentation they used the mean as a representation of a single observation and gave the SD as a measure of the variability in single observations. They also provide the

median, a nonparametric measure of the central value, and the lower and upper quartiles of the variable's distribution, a nonparametric indication of variability. By providing all of this information, they allow the reader to make his or her own assessment of the variability of the data, which does not have to depend on an assumption that the data have a Gaussian (normal) distribution.

Adams et al⁶ and Eggesbe et al⁸ both provide CIs for their measures of variability. The presentation of Adams et al⁶ is preferable because they specify that they are using 95% CIs, whereas Eggesbe et al⁸ only use the term "confidence interval." It is a good assumption that what is intended is a 95% interval, but it is better to state that explicitly.

6. *Incomplete description of the statistical test used.* When describing the statistical tests you have used, think in terms of providing the reader with enough information so that if they had the data, they would be able to do exactly the same test that you did and hopefully get the same result. It is not enough to say you have done a *t* test when in fact you have done a 2-sample *t* test with unequal variances.

In the final paragraph of their "Statistical methods" section, Strunk et al⁴ provide additional information that would allow the reader to duplicate their results. They have also provided information on the statistical program (software) they used for their analysis. This might also be needed for a reader to reproduce results because software programs might differ in how they do calculations, and different programs might give you slightly different results.

A complete cohort of randomly assigned CLIC subjects was used for the model building. The cohort was revised according to variables in the final model to include all possible patients. All summary statistics and analyses were performed by means of SAS Version 8 statistical software

(SAS Institute, Cary, NC). Significance was established at $P < .05$, 2-tailed.

Sorensen et al⁷ also provide information on the analysis package used and, in the section below, additional information that a reader might need to reproduce their results. "All reported P values in this article are based on 2-tailed tests, unless otherwise mentioned. It should be noted that the use of a 2-tailed test when the hypothesis involves an expected increase (or decrease) yields a conservative P value."

RECOMMENDATIONS ON WRITING A STATISTICAL METHODS SECTION

It is important to have the right perspective about correctly reporting statistical tests in the medical literature. It is very easy to think that because misreporting has been going on for so long without seeming to harm medical research, this is not a big issue. In the words of John Bailar,¹¹ however, "... there may be greater danger to the public welfare from statistical dishonesty than from almost any other form of dishonesty." This article is not about dishonesty but about carelessness in the reporting of statistical methods. The harmful results, however, could be the same, and the first step in separating the one from the other is to agree on rigorous reporting in medical journals of statistical methods. This will make it easier for researchers to understand what constitutes the standard of excellence to which all of us should aspire.

Dr Bailar's quote is particularly appropriate because the recommendations for statistical reporting that he and Dr Mosteller laid out in 1988¹² are still the basis for reporting statistical methods that are recommended by the International Committee of Medical Journal Editors. Their 15 points are listed below. More detail can be obtained from the cited article.

1. Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results.
2. When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (eg, confidence intervals).
3. Avoid sole reliance on statistical hypothesis testing, such as the use of P values, which fails to convey important quantitative information.
4. Discuss eligibility of experimental subjects.
5. Give details about randomization.
6. Describe the methods for, and success of, any blinding of observations.
7. Report treatment complications.
8. Give numbers of observations.
9. Report losses to observation (eg, dropouts from a clinical trial).
10. References for study design and statistical methods should be to standard works (with pages stated) when possible rather than to papers where designs or methods were originally reported.

11. Specify any general use computer programs used.
12. Put general descriptions of statistical methods in the Methods section. When data are summarized in the Results section, specify the statistical methods used to analyze them.
13. Restrict tables and figures to those needed to explain the argument of the paper and to assess its support. Use graphs as an alternative to tables with many entries; do not duplicate data in graphs and tables.
14. Avoid nontechnical uses of technical terms in statistics, such as "random" (which implies a randomizing device), "normal," "significant," "correlation," and "ample."
15. Define statistical terms, abbreviations, and most symbols.

Recently, several major physiology journals have adopted guidelines for reporting statistical information that adopt many of the above principles.¹³ These guidelines contain practical and straightforward suggestions for reporting statistical information that should help you to avoid the problems discussed above. The single most important thing to do is to consult a statistician well before you undertake the analysis of your data. If the early steps of design and data collection are done correctly, the actual analysis of the data is usually a relatively minor issue.

REFERENCES

1. Olsen CH. Review of the use of statistics in infection and immunity. *Infect Immun* 2003;71:6689-92.
2. Garcia-Berthou E, Alcaraz C. Incongruence between test statistics and P values in medical papers. *BMC Medical Research Methodology* 2004. Available at: <http://www.biomedcentral.com/1471-2288/4/13>. Accessed in late June 2004.
3. White S. Statistical errors in papers in the British Journal of Psychiatry. *Br J Psychiatry* 1979;135:336-42.
4. Strunk RC, Szefer SJ, Phillips BR, Zeiger RS, Chinchilli VM, Larsen G, et al. Relationship of exhaled nitric oxide to clinical and inflammatory markers of persistent asthma in children. *J Allergy Clin Immunol* 2003;112:883-92.
5. Asero R, Mistrello G, Roncarlo D, Amato S, Zannoni D, Barocci F, et al. Detection of clinical markers of sensitization to profiling in patients allergic to plant-derived foods. *J Allergy Clin Immunol* 2003;112(2):427-31.
6. Adams RJ, Weiss ST, Fuhlbrigge A. How and by whom care is delivered influences anti-inflammatory use in asthma: results of a national population survey. *J Allergy Clin Immunol* 2003;112:445-50.
7. Sorensen B, Streib JE, Strand M, Make B, Giclas PC, Fleshner M, et al. Complement activation in a model of chronic fatigue syndrome. *J Allergy Clin Immunol* 2003;112:397-403.
8. Eggesbe M, Botten G, Stigum H, Nafstad P, Magnus P. Is delivery by cesarean section a risk factor for food allergy? *J Allergy Clin Immunol* 2003;112:420-6.
9. International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to biomedical journals. *Ann Intern Med* 1997;126:36-47.
10. Curran-Everett D. Multiple comparisons: philosophies and illustrations. *Am J Physiol Regul Integr Comp Physiol* 2000;279:R1-8.
11. Bailar JC. Bailar's laws of data analysis. *Clin Pharmacol Ther* 1976;20:113-20.
12. Bailar JC, Mosteller F. Guidelines for statistical reporting in articles for medical journals. *Ann Intern Med* 1988;108:266-73.
13. Curran-Everett D, Benos DJ. Guidelines for reporting statistics in journals published by the American Physiological Society. *Am J Physiol Cell Physiol* 2004;287:C243-5.