

---

## CME review article

This feature is supported by an unrestricted educational grant from AstraZeneca LP

# The reading, writing, and arithmetic of the medical literature, part 2: critical evaluation of statistical reporting

Lyndon Mansfield, MD

---

**Objective:** To offer suggestions to help improve clinicians' understanding of the statistical analyses in the literature and their use of these methods in their own medical writings.

**Data Sources:** Literature searches began at the National Library of Medicine's online database and were traced to primary sources.

**Study Selection:** All referenced information in this article was cited from primary sources.

**Results:** Physicians should be able to determine the variables studied and how they were measured, the comparisons that were made, the difference (with 95% confidence interval) between the groups, the exact *P* value for the difference, the statistical test used in the analysis, whether the data conformed to the assumptions of the test, whether the study had adequate statistical power, and the clinical importance of the difference.

**Conclusion:** Clinicians should know how to interpret statistical results so that they can use medical science to its full extent in treating patients.

*Ann Allergy Asthma Immunol.* 2005;95:315–322.

**Off-label disclosure:** Dr Mansfield has indicated that this article does not include the discussion of unapproved/investigative use of a commercial product/device.

**Financial disclosure:** Dr Mansfield has indicated that in the last 12 months he has not had any financial relationship, affiliation, or arrangement with any corporate sponsors or commercial entities that provide financial support, education grants, honoraria, or research support or involvement as a consultant, speaker's bureau member, or major stock shareholder whose products are prominently featured either in this article or with the groups who provide general financial support for this CME program.

### Instructions for CME credit

1. Read the CME review article in this issue carefully and complete the activity by answering the self-assessment examination questions on the form on page 2.
2. To receive CME credit, complete the entire form and submit it to the ACAAI office within 1 year after receipt of this issue of the *Annals*.

## INTRODUCTION

*"The problem of statistical errors in the medical literature is long-standing, wide-spread, potentially serious, relatively unknown, and not well addressed, despite the fact that most errors occur in the more common applications of statistics."*<sup>1</sup>

The use of hypothesis testing first appeared in the medical literature in the 1930s.<sup>2</sup> Since then, dozens of researchers across a broad range of biomedical disciplines have found that numerous medical articles contain major statistical errors.<sup>1,3–7</sup> The persistence of this situation for more than 70 years reflects the unfortunate reality that few clinicians have a working knowledge of statistics. This study offers sugges-

tions to help improve clinicians' understanding of the statistical analyses in the literature and their use of these methods in their own medical writings. Literature searches began at the National Library of Medicine's online database and were traced to primary sources. All referenced information in this article was cited from primary sources. The questions addressed herein should help readers identify major statistical errors and understand their implications when interpreting statistical results.

## WHAT VARIABLES WERE STUDIED AND HOW WERE THEY MEASURED?

Science is measurement. Thus, the "who, what, when, where, and why" of the measurements must be clearly stated and appropriate for the study. Assuming that you know how something was measured is a common error. Was mucous discharge measured as present or absent? Or as absent, mild,

---

Western Sky Medical Research and Department of Pediatrics, Texas Tech Regional Health Science Center, El Paso, Texas.

Received for publication October 8, 2004.

Accepted for publication in revised form March 10, 2005.

---

moderate, or severe? Or as indicated on a scale of 0 to 100? Was its presence determined by patient self-report, physical examination, or tissue counts?

Once you have ascertained how the variables were measured, you must determine whether the measurements are appropriate for the study. One definition of a current smoker, for example, is anyone who has smoked 1 or more cigarettes in the past 30 days. This definition may make sense in a study of adult patients at a Veterans Affairs hospital but not in a study of adolescents, whose experimentation may not exceed 1 cigarette. Likewise, infection can be defined in different ways. Was the virus isolated from lavage fluid on at least 1 of 6 posttreatment days, or were serum antibody titers elevated by a certain factor after inoculation?

Sometimes surrogate end points are used to measure outcomes because these intermediate points are more easily measured than the clinical end point of interest. For example, the number of rescue uses of an inhaled bronchodilator by asthmatic patients has been used as a surrogate end point when pulmonary function cannot be measured directly. However, studies have shown discordance between pulmonary functions and medication use, and numerous studies have shown a lack of correlation between patients' perceptions of their lung status and objective pulmonary measurements. Bronchodilator use is driven by a patient's subjective perception, and it is not a good surrogate marker for pulmonary function measurements. It may be a good surrogate for control of clinical symptoms. Medication may be an outcome measure. Thus, you need to decide whether the surrogate end point is credible enough to support the study's conclusions.

Composite end points are scores created by combining the results of several components. For example, respiratory function is often measured by the Total Symptom Score, which consists of the sum of the ratings of 4 different respiratory symptoms on a scale of 0 to 4. Thus, the Total Symptom Score could range from 0 (no symptoms) to 16 (maximum symptoms). The use of composite end points raises 2 important questions: What portion of the total score does each component contribute? Are they all likely to be affected equally by the intervention?

#### WHAT COMPARISONS WERE MADE?

Once you know the variables and how they were measured, you need to identify the primary comparison, which forms the basis for the study's design. Although the importance of reporting the primary comparison seems obvious, many studies include dozens of comparisons, some more important than others. With so many comparisons to choose from, authors may report only those that are statistically significant, rather than those they set out to study (a practice called data dredging). In addition, as more comparisons are analyzed, the likelihood increases that a statistically significant *P* value will be interpreted incorrectly to mean that the difference between groups was caused by the intervention.

Often, the primary comparison is the difference between 2 forms of treatment (for example, active drug vs placebo) on

a primary outcome variable. This comparison should be specified at the beginning of the study, and it should be the first one reported in the "Results" section and considered in the "Discussion" section. If it is not, the reason may be that the primary comparison was not statistically significant but a secondary comparison was, and the researchers inappropriately elevated the secondary comparison to a position of primary importance. In a randomized trial, the primary comparison is the one that drives the power analysis: the sample size of the study is based on the size of the difference to be detected in the primary comparison.

Comparisons made after looking at the data—post hoc analyses, unplanned secondary or subgroup analyses, or hypothesis-generating studies—should be labeled accordingly. These comparisons can produce additional insight, but because they were suggested after the data were collected, they should not be presented as hypotheses that were tested by the study.<sup>8</sup> As one researcher stated, "Hypothesis-generating studies (sometimes referred to somewhat contemptuously as 'fishing expeditions') should be identified as such. If the fishing expedition catches a boot, the fishermen should throw it back, not claim that they were fishing for boots."<sup>9</sup>

#### WHAT IS THE DIFFERENCE BETWEEN THE GROUPS?

Typically, a study evaluates the difference between treatment groups on the primary end point. Thus, the difference between group means, medians, or proportions on the primary outcome is the important result. However, this difference can be reported in several ways, leading to different impressions of the outcome (Table 1).<sup>10</sup> For example, in a study comparing levalbuterol to racemic albuterol, levalbuterol can be described as providing a 9% *absolute* risk reduction in hospitalization in asthmatic patients or as a 20% *relative* risk reduction. Both presentations are correct and acceptable, but each imparts a different impression of the reduced risk. For this reason, additional information such as the proportion of hospitalizations for each group and the actual difference between these proportions should be provided.<sup>1</sup> From these data, the other results can be calculated.

#### WHAT IS THE 95% CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN THE GROUPS?

The difference between group outcomes in a study is an estimate of what we would expect to find if that study were repeated in the larger population from which the sample was drawn. We are not as focused on the patients enrolled in the study as we are on the larger issue of how well the treatment will work in the general patient population.

The precision of the estimate is important. Estimates must be precise enough for clinicians to have confidence in basing therapeutic decisions on them. In medicine, the most common measure of precision for an estimate is the 95% confidence interval. A 95% confidence interval indicates the range of values in which we would expect to find our estimated value in 95 of 100 similar studies.

Table 1. Hospitalization Outcomes of 547 Asthma Attacks Among 482 Children Treated With Racemic Albuterol or Levalbuterol in a Randomized, Double-blind, Controlled Trial\*

Outcome	No. (%) of patients		
	Racemic albuterol	Levalbuterol	Total
Hospitalized	123 (45)	99 (36)	222 (41)
Not hospitalized	150 (55)	175 (64)	325 (59)
Total	273 (100)	274 (100)	547 (100)
Eight ways to report the above results			
Absolute risk	The risk (probability) of hospitalization is 45% with racemic albuterol but only 36% with levalbuterol.		
Absolute risk reduction	Levalbuterol reduces the absolute risk of hospitalization by 9% compared with racemic albuterol ( $0.45 - 0.36 = 0.09$ ).		
Relative risk	The risk of hospitalization with use of levalbuterol is 80% of the risk with racemic albuterol ( $0.36/0.45 = 0.8$ ).		
Relative risk reduction	Levalbuterol provides a relative risk reduction in hospitalization of 20% over racemic albuterol ( $0.45 - 0.36/0.45 = 0.2$ ).		
Odds	The odds of hospitalization (compared with not being hospitalized) are 0.82 with racemic albuterol (123/150) but only 0.57 with levalbuterol (99/175).		
Odds ratio	The odds ratio of hospitalization with levalbuterol compared with racemic albuterol is 0.69 ( $0.57/0.82$ ).		
Natural frequencies	Of every 100 asthma attacks treated with racemic albuterol, 45 will result in hospitalization; of every 100 treated with levalbuterol, only 36 will result in hospitalization.		
Number needed to treat	Approximately 11 asthma attacks will need to be treated with levalbuterol for each hospitalization prevented ( $1/0.09$ ).		

\* Adapted from Carl et al.<sup>10</sup>

In the albuterol example cited earlier, the difference between groups (the estimate) is the 9% difference in hospitalization rates. If the 95% confidence interval ranged from 1% to 17%, we would expect the actual difference in hospitalization rates to fall within this range in 95 of 100 similar trials. If even a 1% difference in hospitalization rates is clinically important (the low end of the interval), the confidence interval is said to be homogeneous, and we have more confidence in the efficacy of the drug.

Suppose, however, that we need at least a 5% difference in hospitalization rates to justify changing from one drug to the other. Now the 1% to 17% confidence interval contains clinically unimportant values (those below 5%), and it is said to be heterogeneous. The actual difference in hospitalization rates may be between 1% and 5%, which is below the threshold determined to be clinically important. It may also be above 5%, but we cannot be sure. We want to be 95% confident that we have got it right, which occurs only if the 95% confidence interval contains only clinically important values.

One solution is to increase the sample size, which would provide a more precise estimate. With a larger sample size, if the new confidence interval is determined to be 6% to 12%, for example, we can be 95% confident that the true difference in hospitalization rates will be between 6% and 12%. That range is entirely above the clinically important threshold of 5%.

Confidence intervals are useful because they are expressed in the same units as the estimate, are sensitive to sample size,

and avoid some of the problems encountered when interpreting *P* values. In this example, the 9% difference in hospitalization rates happens to be statistically significant ( $P = .02$ ), and it is significant whether the 95% confidence interval is 1% to 17% or 6% to 12%. The traditional, probabilistic interpretation holds that because chance is not a likely explanation for this 9% difference, we can attribute the difference to the treatment. By providing confidence intervals, however, we emphasize both the size and the clinical relevance of this effect. In so doing, we can help prevent the most common statistical error in the medical literature: confusing statistical significance with clinical importance.

### WHAT IS THE EXACT *P* VALUE FOR THE DIFFERENCE?

Although confidence intervals are slowly becoming the preferred way to report study results, *P* values continue to be used, often in conjunction with confidence intervals. Probability, or *P*, values are a measure of chance as an explanation for a given difference. The lower the *P* value, the less evidence there is to support chance as an explanation for an outcome. If chance is not a likely explanation for the difference (usually, if  $P < .05$ ), we attribute the difference to the effects of the intervention.

Remember that *P* values have no biological interpretation, although they may indicate biological relationships. In classic hypothesis testing, the *P* values calculated from the data are compared with an  $\alpha$  level, which is usually set at .05 or .01.

---

By definition,  $P$  values below the  $\alpha$  level are statistically significant; those above the  $\alpha$  level are not.

One problem with interpreting  $P$  values is the arbitrariness of the  $\alpha$  level. When  $\alpha$  equals .05, for example, a  $P$  value of .051 should be interpreted similarly to a value of .049, but the second is considered statistically significant, and the first is not. For this reason, the exact  $P$  value is preferred to a threshold value (that is,  $P = .02$  is preferred to  $P < .05$ ). The  $\alpha$  level of .05, though commonly used, is also arbitrary; any number between 0 and 1 could be used.

Some authors report that their  $P$  values “trended toward” or “approached” significance. The implication here is that “we almost made it to statistical significance.” In fact,  $P$  values do not “trend” or “approach” anything; they are either significant or not, depending on their relationship to the  $\alpha$  level. (Curiously,  $P$  values never seem to trend away from significance.) For all of these reasons, 95% confidence intervals are now preferred to  $P$  values for reporting results.

### WHAT STATISTICAL TEST WAS USED IN THE ANALYSIS?

$P$  values are calculated using statistical tests. Of the hundreds of statistical tests and procedures available, only a few dozen are widely used (Table 2).<sup>11</sup> Because several tests may be appropriate for any given comparison, the actual test used in making each comparison should be identified. Often, this information will appear in the “Statistical Methods” section at the end of the “Materials and Methods” section, although it may also appear in the “Results” section. A test that is not commonly used should be accompanied by a reference or an explanation of the calculation.

Statistical tests can be 1-tailed or 2-tailed. When either only positive or only negative differences are of interest, a directional hypothesis—or 1-tailed test—is used. Thus, in a study of growth hormone, we are interested only in the probability that the treatment group will be larger than the control group; we do not expect growth hormone to shrink patients in the treatment group. Therefore, only large differences that favor the treatment group can result in a statistically significant  $P$  value. When the direction of the difference is unknown, however, a 2-tailed test is preferred. Large differences between means that favor either the treatment group or control group can result in a significant  $P$  value.

On occasion, a difference that is statistically significant using a 1-tailed test is not significant using a 2-tailed test. When  $P$  values were more commonly used to interpret results (that is, before 95% confidence intervals became popular), this difference was sometimes important. Two-tailed tests are more conservative and more often used, although 1-tailed tests can be used if justified by the researchers.

Finally, the statistical software package used in the analysis should be identified. Commercially available packages, which have usually been validated over time and experience, are preferred over custom packages. In addition, different packages may use different algorithms in their calculations, which can affect results.

### DID THE DATA CONFORM TO THE ASSUMPTIONS OF THE TEST?

All statistical tests are based on assumptions. These assumptions, if violated, can reduce confidence in the results. Therefore, studies should include a statement confirming that the assumptions of the statistical tests were met by the data, although such statements are rarely found in the literature.

The most common assumption is that data from the groups being compared are normally distributed, allowing them to be analyzed using a class of tests called parametric tests. Data that are markedly skewed, however, violate the assumptions of parametric tests, and the results may not be valid. In such cases, the data may be mathematically transformed to a more normal distribution; if so, the success of the transformation should be reported. A more common solution, however, is to use a nonparametric test, which makes no assumptions about the shape of the distribution of the data. For example, the  $t$  test is a parametric test; a common nonparametric equivalent is the Wilcoxon rank sum test (also called the Mann-Whitney  $U$  test).

As a reader, you may be able to detect 2 inconsistencies in the application of parametric and nonparametric tests. Data are generally—if inappropriately—described using the mean and SD. However, the mean and SD should be used *only* to describe normally distributed data.<sup>1</sup> (Remember that 1 SD on either side of the mean value indicates the range that contains approximately 68% of the data.) Thus, the first inconsistency is that if the SD is greater than half the mean (and negative values are not possible), the data are likely skewed enough to warrant a nonparametric test. Skewed data should usually be described using the median and range or interquartile range (the range of values that encompasses the middle 50% of the data).

Remember that the mean and SD are associated with normally distributed data and parametric tests, whereas the median and interquartile range are associated with skewed data and nonparametric tests. Thus, the second inconsistency is apparent when data that are presented as means and SDs are analyzed with a nonparametric test. In such cases, although the data may have been analyzed correctly, they probably have been reported incorrectly.

Another common statistical error is analyzing paired data with unpaired tests. Paired data are data collected from the same subject, for example, before and after an intervention, or data that are collected from subjects who have been matched on selected characteristics. Paired tests take this relationship into account by keeping track of the within-pair changes, as well as the group means (Fig 1). Independent tests make no such assumption.

Another assumption that is often violated occurs in linear regression analysis. A least-squares regression line can be calculated to fit any scatterplot, but the line will predict well only if the variables being compared are, in fact, linearly related (Fig 2). To determine whether a relationship is linear, an analysis of residuals is usually conducted. In this analysis, the differences

Table 2. Definitions of Some Common Statistical Terms and Tests\*

Term or test	Definition
Levels of measurement	A classification of how much information is included in a measurement. From least to most, the levels are <i>nominal</i> (named categories), <i>ordinal</i> (ranked categories), <i>discrete</i> (counts on a scale of equal intervals of whole numbers), and <i>continuous</i> (measurements on a scale of equal intervals and that can include fractions).
<i>P</i> value	A probability value. Indicates the likelihood that the difference found in the study could be the result of chance, under the assumption that the intervention had absolutely no effect (the null hypothesis).
Alpha ( $\alpha$ ) level or alpha error	The threshold that defines statistical significance. <i>P</i> values from the study that are smaller than the $\alpha$ level are, by definition, statistically significant. Thus, results do not “trend toward significance,” they either are or are not significant, depending on their relationship to the $\alpha$ level. Technically, the probability of committing a type I error: attributing a difference to an intervention when chance is the more likely explanation.
Beta ( $\beta$ ) level or beta error or statistical power ( $1 - \beta$ )	The ability of a study to detect a difference of a given size if such a difference actually exists in the data. Related to sample size; larger studies have more statistical power. Technically, the probability of committing a type II error: attributing a lack of an effect to a weak intervention when the lack of sufficient data is the more likely explanation.
Confidence interval (CI)	A measure of the precision of an estimate. Most research results are estimates and should thus be reported with CIs. A 95% CI identifies the range of values in which the result would be expected to occur in 95 of 100 similar studies on the same population.
<i>t</i> test	A parametric hypothesis test used to compare the mean values of 2 groups of normally distributed continuous data.
Wilcoxon rank sum test (also called the Mann-Whitney <i>U</i> test)	A hypothesis test used to compare the median values of 2 groups of non-normally distributed (skewed) continuous data; a nonparametric equivalent of the <i>t</i> test.
Analysis of variance (ANOVA)	A parametric hypothesis test used to compare the mean values of 3 or more groups of normally distributed, continuous data.
Kruskal-Wallis test	A hypothesis test used to compare the median values of 3 groups of non-normally distributed (skewed) continuous data; a nonparametric equivalent of ANOVA.
Correlation coefficient	A number ranging from $-1.0$ to $+1.0$ , indicating whether and how a change in 1 variable is related to a change in another. Different correlation coefficients are used for variables with different levels of measurement: <ul style="list-style-type: none"> <li>• Pearson product-moment correlation coefficient (<i>r</i>): for 2 normally distributed, continuous variables</li> <li>• Spearman rank correlation coefficient (<math>\rho</math>): for 2 continuous variables of any distribution</li> <li>• Kendall rank correlation coefficient (<math>\tau</math>): for 2 ordinal variables or 1 ordinal and 1 continuous variable</li> <li>• Point biserial correlation coefficient: for a continuous variable and a categorical variable with 2 levels (recovery status: recovered or not)</li> <li>• Point multiserial correlation coefficient: for a continuous variable and a categorical variable with 3 or more levels (disease severity: mild, moderate, severe)</li> <li>• Intraclass and interclass correlation coefficients: assess agreement within and between observers, respectively</li> </ul>
Regression analysis	A set of statistical procedures for predicting the value of 1 variable from known values of 1 (simple regression) or more (multiple regression) variables. Analyses (or models) predicting a dichotomous variable are usually logistic regression models; those predicting a continuous variable are often linear regression models.
Beta ( $\beta$ ) weight	A regression coefficient; tells how much 1 variable will change for each unit change in the other.
Chi-square ( $\chi^2$ ) test	A set of hypothesis tests for comparing nominal or categorical data: <ul style="list-style-type: none"> <li>• <math>\chi^2</math> test for independence (or test of association or Pearson <math>\chi^2</math> test)</li> <li>• <math>\chi^2</math> test for proportions: not a test of association. Determines whether the difference between group proportions differs significantly from zero (a test of association assess the mix of frequencies in a contingency table)</li> <li>• <math>\chi^2</math> test for goodness-of-fit (to a known distribution of proportions)</li> <li>• Exact tests for small samples (eg, Fisher exact test)</li> </ul>

\* From Lang.<sup>11</sup>

between each measured value and the predicted value of the response variable are graphed. When the graph shows a narrow band of values close to a difference of 0, the relationship is

probably linear (Fig 3). Any other pattern usually indicates a nonlinear relationship. Authors should report whether and how they evaluated the relationship for linearity.

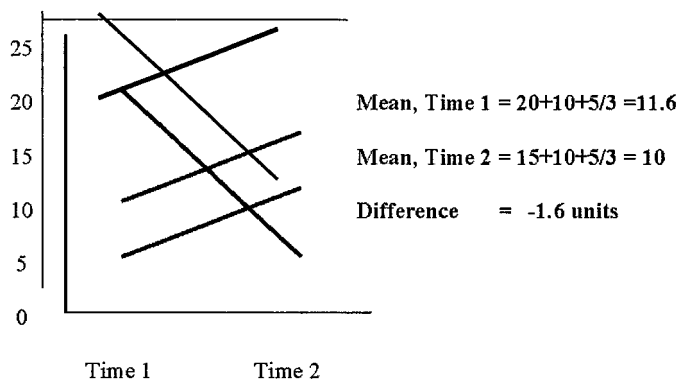


Figure 1. Paired statistical tests keep track of within-pair changes and group means. Here, the within-pair changes are more pronounced than would be apparent by simply looking at the 1.6-unit difference in means. An unpaired test is used when the groups being compared are independent of one another, such as comparing a mean value in a group of men with that of a group of women.

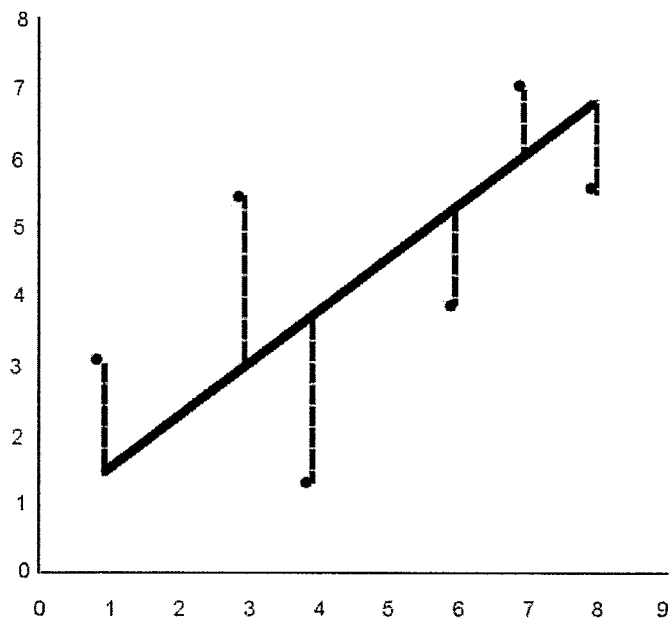


Figure 2. A residual is the difference between the observed value of Y (the dot representing a data point) and the value of Y predicted by the regression line for a given value of X (the Y value where the line crosses the value of X).

An example of the need for caution in interpreting linear regression analysis can be found in a study that examined the effects of long-term inhaled triamcinolone therapy on bone density in premenopausal women.<sup>12</sup> Based on regression analysis, the researchers calculated a significant bone loss of 0.00044 g/cm<sup>2</sup> per puff of triamcinolone per year on both the total hip and the trochanter ( $P = .01$  and  $P = .005$ , respectively). They found no dose effect on the femoral neck or spine. Although the study lasted 3 years, the authors projected

their findings to 20 years, suggesting that women in the study group could have more than twice the risk of hip fracture in the future compared with untreated women.

Scattergrams of the subjects' data were included in the results of the study. Analysis of these scattergrams reveals the clinically significant observation that many subjects, across a wide range of puffs per day, showed increased bone density. More specifically, in the group that inhaled the greatest number of daily puffs, 10 subjects showed increased density, and no subject had the bone density loss predicted by the authors. It is difficult to discern the clinically important steroid effect suggested by the authors based on the scatter of all the data. In addition, because no control, non-steroid-exposed population was included in the study, we cannot determine whether the scatter is normal or abnormal.

### DID THE STUDY HAVE ADEQUATE STATISTICAL POWER?

A study that reports a significant  $P$  value for the primary comparison is often referred to as positive. Otherwise, it is referred to as negative. Negative studies are often incorrectly interpreted to mean that the intervention was ineffective. A study can be negative either because the intervention was ineffective or because not enough data were collected. Only studies with adequate statistical power can rule out insufficient data as an explanation for lack of difference. In fact, underpowered, negative studies are not negative at all; rather, they are inconclusive. Statistician Frederick Mosteller illustrates the problem of low statistical power by adding in italics what most authors should say but do not: "The increase in infection rate using the new methods was not statistically significant. . . and there was not 1 chance in 10 that we would have detected a 30% increase in rate."<sup>13</sup>

Adequate statistical power usually means that enough patients have been enrolled in the study to find a given difference, if such a difference exists. The number of patients required is determined by a power calculation, which incorporates several factors, such as the minimum size of the difference to be found, the variability of the data, the  $\alpha$  level, and the desired statistical power. Statistical powers of 80% or 90% are typical, meaning that the study has an 80% or 90% chance to detect a specified difference in the primary end point if such a difference really exists.

Statistical power is even more important in equivalence or noninferiority trials, in which a study is designed to show that one drug performs just as well as another. In these trials, the results cannot differ greatly, which means that larger samples are needed to rule out all but the smallest differences. Positive studies find statistically significant differences and so, by definition, are adequately powered to detect those differences.

### WHAT IS THE CLINICAL IMPORTANCE OF THE DIFFERENCE?

The most common error in interpreting statistics, if not the most serious, is to confuse statistical significance with clin-

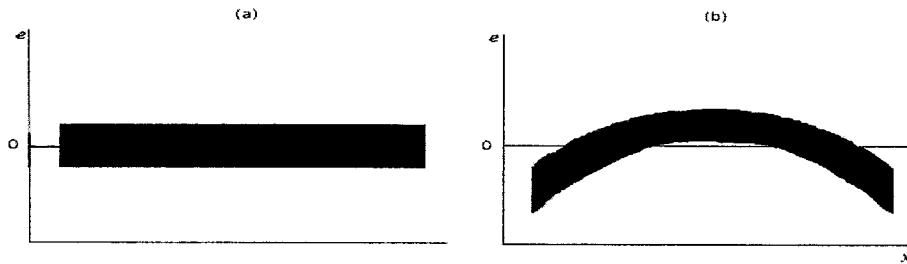


Figure 3. When the graphed residuals remain close to 0 over the range of values, the regression line accurately represents the linear relationship of the data (a). Any other pattern (b) indicates that the relationship is not linear, which may mean that linear regression analysis should not be used.

ical importance.<sup>1</sup> As described earlier, *P* values have no clinical interpretation. Clinical importance must be determined by the size of the difference or, more precisely, by the estimated size of the difference and its measure of precision (the 95% confidence interval), as well as by other human and medical considerations.

The inclusion of data in a study does not mean that such data are clinically meaningful. One study that evaluated montelukast, loratadine, and the combined drugs vs placebo in the treatment of seasonal allergic rhinitis presented Rhinitis Quality of Life Questionnaire (RQLQ) data. All 3 treatment arms showed statistically significant improvement in RQLQ vs placebo, with improvement in scores ranging from 0.25 to 0.35. However, the minimally important *clinical* difference on RQLQ is 0.5, so these results appear to have no clinical meaning. In this example, there is a valid standard for a clinically meaningful minimal difference. This is not always the case, and readers must be prepared to interpret findings based on their own clinical judgment.<sup>14</sup>

## CONCLUSION

Statistical analyses are an integral part of medical science, but they are not akin to practicing medicine. Statistics are based on groups, whereas medicine is practiced on individuals. Statistics require adequate amounts of data, whereas medical decisions must often be made with insufficient data. Statistics also require measurement, whereas medicine sometimes requires intuition. Nevertheless, clinicians should know how to interpret statistical results so that they can use medical science to its full extent in treating patients. The third, and concluding, article in this series describes an approach for the critical appraisal of biomedical research.

## ACKNOWLEDGMENTS

The author would like to thank Sanofi-Aventis for its generous support to the editorial assistance and development of this article. I also thank Tom Lang for his efforts and assistance in preparation of this article.

## REFERENCES

1. Lang T, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. Philadelphia, PA: American College of Physicians; 1997.
2. Morgan P. Confidence intervals: from statistical significance to clinical significance. *CMAJ*. 1989;141:881–883.
3. Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation*. 1980;61:1–7.
4. Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA*. 1966;195:1123–1128.
5. White SJ. Statistical errors in papers in the *British Journal of Psychiatry*. *Br J Psychiatry*. 1979;135:336–342.
6. Hemminki E. Quality of reports of clinical trials submitted by the drug industry to the Finnish and Swedish control authorities. *Eur J Clin Pharmacol*. 1981;19:157–165.
7. Gore SM, Jones G, Thompson SG. *The Lancet's* statistical review process: areas for improvement by authors. *Lancet*. 1992;340:100–102.
8. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. *Ann Intern Med*. 1992;116:78–84.
9. Mills JL. Data torturing [letter]. *N Engl J Med*. 1993;329:1196–1199.
10. Carl JC, Myers TR, Kirchner HL, Kerckmar CM. Comparison of racemic albuterol and levalbuterol for treatment of acute asthma. *J Pediatr*. 2003;143:731–736.
11. Lang T. Interpreting and reporting public health and medical research: techniques and 13 key questions. Continuing education course handout presented at: Annual Meeting of the American Public Health Association; November 15–19, 2003; San Francisco, CA.
12. Israel E, Banerjee TR, Fitzmaurice GM, Kotlov TV, LaHive K, LeBoff MS. Effects of inhaled glucocorticoids on bone density in premenopausal women. *N Engl J Med*. 2001;345:941–947.
13. Mosteller F, Gilbert JP, McPeck B. Reporting standards and research strategies for controlled trials. *Control Clin Trials*. 1980;1:37–58.
14. Nayak AS, Philip G, Lu S, et al. Efficacy and tolerability of montelukast alone or in combination with loratadine in seasonal allergic rhinitis: a multicenter, randomized, double-blind, placebo-controlled trial performed in the fall. *Ann Allergy Asthma Immunol*. 2002;88:592–600.

Requests for reprints should be addressed to:  
 Lyndon Mansfield, MD  
 1901 Arizona Ave  
 El Paso, TX 79902  
 E-mail: immunman@pol.net

---

---

**Objectives:** After reading this article, participants should be able to demonstrate an increased understanding of their knowledge of allergy/asthma/immunology clinical treatment and how this new information can be applied to their own practices.

**Participants:** This program is designed for physicians who are involved in providing patient care and who wish to advance their current knowledge in the field of allergy/asthma/immunology.

**Credits:** ACAAI designates each Annals CME Review Article for a maximum of 2 category 1 credits toward the AMA Physician's Recognition Award. Each physician should claim only those credits that he/she actually spent in the activity. The American College of Allergy, Asthma and Immunology is accredited by the Accreditation Council for Continuing Medical Education to sponsor continuing medical education for physicians.

---

### CME Examination

1–5, Mansfield L. 2005;95:315–322.

### CME Test Questions

1. A problem of composite scores is
  - a. they do not tell the reader the proportion contributed by each component of the composite
  - b. they have no clinical meaning
  - c. they cannot be analyzed with statistical methods
2. In a randomized placebo-controlled trial
  - a. an active treatment is compared with placebo
  - b. the investigators do not know which treatment arm the subject is on
  - c. statistical power analysis is based on the primary efficacy comparison
  - d. a, b, and c
  - e. a and c
3. The 95% confidence interval
  - a. includes the range of values of our estimate for 95 of 100 similar studies
  - b. may contain both clinically meaningful and nonclinically meaningful values
  - c. generally becomes more precise if the sample size is increased
  - d. none of the above
  - e. all of the above
4. The Wilcoxon rank sum test
  - a. involves comparing medians
  - b. is a nonparametric test
  - c. is somewhat comparable to the *t* test
  - d. all of the above
  - e. none of the above
5. When a *P* value of .05 or less is not achieved
  - a. the study samples may be too small to demonstrate a true difference
  - b. the null hypothesis is accepted
  - c. if the statistical power for the comparison was 90% the value is accepted
  - d. a and c
  - e. all of the above

**Answers found on page 380.**