

COMMON STATISTICAL ERRORS EVEN YOU CAN FIND*

PART 1: ERRORS IN DESCRIPTIVE STATISTICS AND IN INTERPRETING PROBABILITY VALUES

Tom Lang, MA
Tom Lang Communications

“Critical reviewers of the biomedical literature have consistently found that about half the articles that used statistical methods did so incorrectly.”¹

Statistical probability was first discussed in the medical literature in the 1930s.² Since then, researchers in several fields of medicine have found high rates of statistical errors in large numbers of scientific articles, even in the best journals.³⁻⁶ The problem of poor statistical reporting is, in fact, longstanding, widespread, potentially serious, and almost unknown, despite the fact that most errors concern basic statistical concepts and can be easily avoided by following a few guidelines.⁷

The problem of poor statistical reporting has received more attention with the rise of the evidence-based medicine movement. Evidence-based medicine depends on the quality of published research; that is, evidence-based medicine is literature-based medicine. As a result, several groups have proposed reporting guidelines for different types of trials,⁸⁻¹⁰ and a comprehensive set of guidelines for reporting statistics in medicine has been compiled from an extensive review of the literature.¹¹

In a series of articles, I will describe several of the more common statistical errors found in the biomedical literature, errors that can be identified even by those who know little about statistics. These guidelines are but the tip of the iceberg; readers who want to know more about the iceberg should consult more detailed texts,¹¹ as well as other references cited in this series.

The field of statistics can be divided into two broad areas: **descriptive statistics**, which is concerned with how to describe samples of data collected in a research study, and **inferential statistics**, which is concerned with how to estimate (or infer) from the sample the characteristics of

the population from which the sample was selected. In this article, I describe errors made in defining variables, in summarizing the data collected about these variables, and in interpreting probability (*P*) values.

Errors in Descriptive Statistics

Error #1: Not Defining Each Variable in Measurable Terms

Science is measurement. Researchers need to tell us what they measured—the variables—and how they measured them, by providing the **operational definition** of each variable. For example, one operational (measurable) definition of hypertension is a systolic blood pressure of 140 mm Hg or higher, and an operational definition of obesity is a body mass index above 27.3 for women and above 27.8 for men.

Variables relating to concepts or behaviors may be more difficult to measure. Depression defined as a score of more than 50 on the Zung Depression Inventory is operationally defined, but how well the Inventory actually measures depression can be debated. In one major U.S. survey, a “current smoker” is anyone who smoked one cigarette in the 30 days before the survey. Although this definition is not an obvious one, it is nevertheless an “operational” one, and we at least know who “current smokers” are in the survey, even if we disagree with the definition.

Error #2: Not Providing the Level of Measurement of Each Variable

Level of measurement refers to how much information is collected about the variable. For practical purposes, there are three levels of measurement: nominal, ordinal, and continuous. At the lowest level are **nominal data**, which consist of two or more nominal, or named, categories that have no inherent order. Blood type defined as type A, B, AB, or O is measured at the nominal level of measurement.

Ordinal data consist of categories that *do* have an inherent order and can be sensibly ranked. A person may

*This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The *AMWA Journal* gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.

be described as short, medium, or tall. We may not know the exact height of the patients studied, but we do know that a person in the tall category is taller than one in the medium category, who, in turn, is taller than one in the short category.

Continuous data consist of values along a continuous measurement scale, such as height measured in centimeters or as blood pressure measured in millimeters of mercury. Continuous data are the highest level of measurement because they tell how far each point value is from any other value on the same scale.

Researchers need to specify the level of measurement for each variable. For example, they may wish to characterize a patient's blood pressure as a nominal variable (either elevated or not elevated), as an ordinal variable (hypotensive, normotensive, or hypertensive), or as a continuous variable (the systolic pressure in millimeters of mercury). The levels of measurement of response and explanatory variables are important because they determine the type of statistical test that can be used to analyze relationships. Different combinations of levels of measurement require different statistical tests.

Error #3: Dividing Continuous Data into Ordinal Categories Without Explaining Why or How the Categories Were Created

To simplify statistical analyses, continuous data, such as height measured in centimeters, are often separated into two or more ordinal categories, such as short, medium, and tall. Reducing the level of measurement in this way also reduces the precision of the measurements, however, as well as reducing the variability in the data. Authors should explain why they chose to lose this precision. In addition, they should explain how the boundaries of the ordinal categories were determined, to avoid the appearance of bias. In some cases, the boundaries (or cut points) that define the categories can be chosen to favor certain results.

Error #4: Using the Mean and Standard Deviation to Describe Continuous Data That Are Not Normally Distributed

Unlike nominal and ordinal data, which are easily summarized as the number or percent of observations in each category, continuous data can be graphed to form distributions. Distributions are usually described with a value summarizing the bulk of the data—the mean, median, or mode—and a range of values that represent the variation of the data around the summary value—the range, the interpercentile range, or the standard deviation (SD).

Normal distributions are appropriately described with any of the above descriptive statistics, although the mean and the SD are used most commonly. In fact, the mean and the SD should be used *only* to describe approximately normal distributions. By definition, about 67% of the values of a normal distribution are within ± 1 SD of the mean, and about 95% are within ± 2 SDs. **Non-normal or skewed distributions**, however, are *not* appropriately described with the mean and the SD. The **median** value (the value that divides observations into an upper and a lower half) and the **interquartile range** (the range of values that include the middle 50% of the observations) are more appropriate for describing non-normally distributed data.

Most biologic data are not normally distributed, so the median and interquartile range should be more common than the mean and the SD. A useful rule of thumb is that if the SD is greater than half of the mean (and negative values are not possible), the data are not normally distributed.

Error #5: Using the Standard Error of the Mean (SEM) As a Descriptive Statistic

Unlike the mean and the SD, which are *descriptive statistics* for a *sample* of (normally distributed) data, the **standard error of the mean (SEM)** is a *measure of precision* for an estimated characteristic of a *population*. (One SEM on either side of the estimate is essentially a 67% confidence interval [see later]. However, the SEM is often reported instead of the SD. The SEM is always smaller than the SD, and so its use makes measurements look more precise than they are. In addition, the preferred measure of precision in the life sciences is the 95% confidence interval. Thus, measurements (when normally distributed) should be described with the mean and SD, not SEM, and an estimate should be accompanied by the 95% confidence interval, not the SEM.

Errors in Interpreting Probability (P) Values

“We think of tests of significance more as methods of reporting than for making decisions because much more must go into making medical policy than the results of a significance test.”¹²

Probability (*P*) values can be thought of as the amount of evidence in favor of chance as the explanation for the difference between groups. When the probability is small, usually less than five times in 100, chance is rejected as the cause, and the difference is attributed to the intervention under study; that is, *P* values indicate mathematical probability, not biologic importance. Probability values are compared to the alpha level that

defines the threshold of statistical significance. Alpha is often set at 0.05. A *P* value below alpha is “statistically significant”; a *P* value above alpha is “not significant at the 0.05 level.” This all-or-none interpretation of a *P* value and the fact that any alpha level is arbitrary are other causes of misinterpretation.

A *P* value can help to decide whether, say, two groups are significantly different. The *lack* of statistical significance, however, does not necessarily mean that the groups are similar. Concluding that groups are equivalent because they do not differ significantly is another common misinterpretation.

Error #6: Reporting Only *P* Values for Results

The problems described have led journals to recommend reporting the 95% confidence interval for the difference between groups (that is, for the “estimate”) instead of, or in addition to, the *P* value for the difference.¹³ The following examples show the usefulness of confidence intervals.¹¹

- *The effect of the drug on lowering diastolic blood pressure was statistically significant ($P < 0.05$).* Here, the *P* value could be 0.049; statistically significant (at the 0.05 level) but so close to 0.05 that it should be interpreted similarly to a *P* value of, say, 0.051, which is *not* statistically significant. In addition, we do not know by how much the drug lowered diastolic pressure; that is, we cannot judge the clinical importance of the reduction.
- *The mean diastolic blood pressure of the treatment group dropped from 110 to 92 mm Hg ($P = 0.02$).* This presentation is the most typical. The values before and after the test are given, but not the difference. The mean drop—the 18-mm Hg difference—is statistically significant, but it is also an *estimate* of the drug’s effectiveness, and without a 95% confidence interval, the precision (and therefore the usefulness) of the estimate cannot be determined.
- *The drug lowered diastolic blood pressure by a mean of 18 mm Hg, from 110 to 92 mm Hg (95% CI = 2 to 34 mm Hg; $P = 0.02$).* The confidence interval indicates that if the drug were to be tested on 100 samples similar to the one reported, the average drop in blood pressure would range between 2 and 34 mm Hg in 95 of the 100 samples. A drop of only 2 mm Hg is not clinically important, but a drop of 34 mm Hg is. So, although the *mean* drop in blood pressure in this particular study was statistically significant, the expected difference in blood pressures may not always be clinically important; that is, the study results are actually *inconclusive*. For conclusive results, more patients probably need to be studied to narrow the

confidence interval until *all* or *none* of the values are clinically important.

Error #7: Not Confirming That the Assumptions of Statistical Tests Were Met

Statistical tests may not give accurate results if their assumptions are violated.¹⁴ For this reason, both the name of the test and a statement that its assumptions were met by the data should be included when reporting statistical analyses. The most common errors are

- Using parametric tests (which require data to be normally distributed) when the data are skewed. In particular, when comparing two groups, the Student *t* test is often used when the Wilcoxon rank sum test (or another nonparametric test that does not assume normally distributed data) is more appropriate.
- Using tests for independent samples on paired samples, which require tests for paired data. Again, the Student *t* test is often used when a paired *t* test is required.
- Using linear regression analysis without establishing that the relationship between variables is, in fact, linear. (The assumption of linearity may be tested by what is called an analysis of “residuals.”¹¹)

Error #8: Interpreting Nonstatistically Significant Results As “Negative” When They Are, in Fact, Inconclusive

A researcher who finds no statistically significant difference between experimental groups must decide whether the lack of difference means that the groups were, in fact, similar (the intervention made no difference), or that too few data were collected to detect a meaningful difference. This decision is usually made with a **power calculation**, which determines how many subjects must be studied to have a given chance of detecting a given difference, if such a difference is there to be detected.

Unfortunately, many studies reporting nonstatistically significant findings are underpowered and, therefore, do not provide conclusive answers.¹⁵ The researchers found no difference, but neither can they rule out the existence of a difference. Absence of proof is not proof of absence.

In inadequately powered studies, statistically insignificant results are truly negative: the groups being compared are, in fact, similar because no difference was found, but a difference *could* have been found had it existed in the data. Adequate power is especially important in equivalence trials (or noninferiority trials), which are conducted to establish that one drug is as good as another.

Error #9: Not Reporting Whether or How Adjustments Were Made for Multiple Hypothesis Tests

Many studies report several *P* values, which increases the risk of making a **type I error**: concluding that the difference found is the result of an intervention when chance is a more likely explanation. For example, to compare each of six groups to all the others, 15 pair-wise statistical tests—15 *P* values—are needed. Without adjusting for these multiple tests, the chance of making a type I error rises from 5 times in 100 (the typical alpha level of 0.05) to 55 times in 100 (an alpha of 0.55).

The multiple testing problem may be encountered when

- **Establishing group equivalence** by testing each of several baseline characteristics for differences between groups (hoping to find none)
- Performing **multiple pair-wise comparisons**, which occurs when three or more groups of data are compared two at a time in separate analyses
- Testing **multiple endpoints** that are influenced by the same set of explanatory variables
- Performing **secondary analyses** of relationships observed during the study but not identified in the original study design
- Performing **subgroup analyses** not planned in the original study
- Performing **interim analyses of accumulating data** (one or more endpoints measured at several different times)
- **Comparing groups at multiple time points** with a series of individual group comparisons (repeated-measures procedures)

Adjusting for multiple comparisons is sometimes optional. However, readers need to know whether or not adjustments were made and, if so, what adjustments were involved.¹⁶ The Bonferroni correction is a common adjustment, for example.

Multiple testing is often needed to explore new relationships among data; however, exploratory analyses should be reported as exploratory. Data dredging—performing *undisclosed* analyses to compute many *P* values to find *something* that is statistically significant (and, therefore, worth reporting)—is poor science.

Error #10: Confusing Statistical Significance with Biologic Importance

As described here, many researchers interpret a statistically significant *P* value as indicating a biologically important result.¹⁷ In fact, *P* values have no biologic interpretation. The nature and size of the difference must be judged to determine biologic importance. Perhaps the best way to remember this most common of statistical errors, as well as to close this article, is with a quote from statistician John Yancy: “It has been said that a fellow with one leg frozen in ice and the other leg in boiling water is comfortable—on average.”¹⁸

References

1. Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation*. 1980;61:1-7.
2. Mainland D. Chance and the blood count. *Can Med Assoc J*. 1934; June:656-658.
3. Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA*. 1966;195:1145.
4. White SJ. Statistical errors in papers in the *Br J Psychiatry*. 1979;135:336-342.
5. 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.
6. Gore SM, Jones G, Thompson SG. The *Lancet's* statistical review process: areas for improvement by authors. *Lancet*. 1992;340:100-102.
7. George SL. Statistics in medical journals: a survey of current policies and proposals for editors. *Med Pediatric Oncol*. 1985;13:109-112.
8. Altman DG, Schulz KF, Moher D, et al, for the CONSORT Group. The CONSORT statement: revised recommendations for improving the quality of parallel-group randomized trials. *Ann Intern Med*. 2001;134:657-62; *Lancet*. 2001;357:1191-1194; *JAMA*. 2001;285:1987-1991.
9. Stroup D, Berlin J, Morton S, et al. Meta-analysis of observational studies in epidemiology [MOOSE]: a proposal for reporting. *JAMA*. 2000;283:2008-2012.
10. Moher D, Cook DJ, Eastwood S, et al, for the QUORUM Group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. *Lancet*. 1999;354:1896-1900.

11. Lang T, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*. Philadelphia: American College of Physicians; 1997.
12. Mosteller F, Gilbert JP, McPeck B. Reporting standards and research strategies for controlled trials. *Control Clin Trials*. 1980;1:37-58.
13. Bailar JC, Mosteller F. Guidelines for statistical reporting in articles for medical journals. *Ann Intern Med*. 1988;108:266-273.
14. DerSimonian R, Charette LJ, McPeck B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med*. 1982;306:1332-1337.
15. Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Control Clin Trials*. 1989;10:31-56.
16. Pocock SJ, Hughes MD, Lee RJ. Statistical problems in the reporting of clinical trials: a survey of three medical journals. *N Engl J Med*. 1987;317:426-432.
17. Ellenbaas RM, Ellenbaas JK, Cuddy PG. Evaluating the medical literature, part II: statistical analysis. *Ann Emerg Med*. 1983;12:610-620.
18. Yancy JM. Ten rules for reading clinical research reports. *Am J Surg*. 1990;159:553-559.

GUIDEBOOK TO BETTER MEDICAL WRITING

by Robert L. Iles

“The best basic manual on medical writing... everything you need to know about developing a clear, persuasive paper that stands a good chance of publication by a peer-reviewed journal.” Barbara G. Cox, MedEdit Associates, Gainesville, FL. (amazon.com book review)

“Iles has succeeded in boiling down the essentials of medical writing into a cogent handbook.” Linda M. Bonnell, PharmD, *AMWA Journal*, 1999;14:31.

“A concise, no-nonsense approach... provides readers with a series of excellent tips...helpful in my own medical writing and consulting service.” Thomas Buckingham, MD, Bratislava, Slovak Republic. (amazon.com book review)

“Although the focus is on clinical articles, what Iles has to say applies to most scientific writing...” Jude Richard, *CBE Views*, 1999;22:201.

Read an excerpt at www.medwriting.com

Send me _____ copy(ies) at \$ **27.95 ea** plus \$3.50 shipping and handling U.S.

25% discount, five or more copies!

Please print

Name _____

Organization _____

Street address _____

City, state, ZIP _____

Enclosed is check money order

Charge to my Visa MasterCard

____ - ____ - ____ - ____

Expiration date _____

Island Press
1065 Wyckford Rd
Olathe, KS 66061
Fax: (913) 782-7138



COMMON STATISTICAL ERRORS EVEN YOU CAN FIND*

PART 2: ERRORS IN MULTIVARIATE ANALYSES AND IN INTERPRETING DIFFERENCES BETWEEN GROUPS

Tom Lang, MA
Tom Lang Communications

This article is the second in a series in which I describe several of the more common statistical errors in the biomedical literature. The first article in the series focused on 10 errors in descriptive statistics and in interpreting probability, or *P* values.¹ Here, I provide an overview of multivariate analyses (regression analysis and analysis of variance, or ANOVA) and describe nine errors in interpreting differences between groups.

An Overview of Multivariate Analyses

The most common forms of multivariate analyses in medicine are regression analysis and ANOVA. The two methods are similar. Both are used in studies involving two or more explanatory variables. In general, ANOVA is used to assess *categorical* explanatory variables, whereas regression analysis is used to assess *continuous* explanatory variables. When a study includes both categorical and continuous and explanatory variables, the analysis may be called either multiple regression analysis or analysis of covariance (ANCOVA). The results of multivariate procedures are referred to as models (equations), because they seek to describe the mathematical relationships among the variables so that one value can be predicted from the others.

The most common types of multiple regression analysis are the following:

- **Linear regression**, in which two or more explanatory variables are used to predict the value of a continuous response variable
- **Logistic regression**, in which two or more explanatory variables are used to predict the value of a binomial response variable (alive or dead, healed or not healed)

*This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The *AMWA Journal* gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.

- **Cox proportional hazards regression**, in which two or more explanatory variables are used to predict the time to an event (such as the time from surgery to death)

The most common ANOVA procedures are one-way ANOVA, two-way ANOVA, multi-way ANOVA, ANCOVA, and repeated-measures ANOVA.² Unfortunately, these procedures take more space to explain.

- **One-way ANOVA** assesses the effect of a *single categorical explanatory variable* (sometimes called a factor) on a single continuous response variable. The factor (category) also has three or more alternatives (or levels or values; for example, the category of blood type has four alternatives: A, B, AB, or O). When there are only two alternatives (two groups), this analysis reduces to the Student *t* test.

Example: Women with osteoporosis have been randomly assigned to one of three groups: a standard treatment, a new treatment, or a placebo. The response variable is the change in bone mineral density, a continuous variable. The explanatory variable is the form of treatment, which distinguishes each group. The results can be analyzed with one-way ANOVA.

- **Two-way ANOVA** assesses the effect of *two categorical explanatory variables* (again, sometimes called factors) on a single continuous response variable.
- **Multi-way ANOVA** assesses the effect of *three or more categorical explanatory variables* (still called factors) on a single continuous response variable.

Example: To the previous example, the addition of more categorical explanatory variables, such as diet

(vegetarian or nonvegetarian) and alcohol consumption (less than 2 ounces of alcohol per day, 2 to 5 ounces per day, or 6 ounces or more per day), would move the analysis from two-way to four-way ANOVA, or simply, multi-way ANOVA.

- **ANCOVA** assesses the effect of one or more categorical explanatory variables *while controlling for the effects of some other (possibly continuous) explanatory variables* (now called covariates) on a single continuous response variable.

Example: To the previous example, we now may wish to control for the severity of disease. Women with more severe osteoporosis may have different bone mineral densities than women with less severe disease. If we are to study the relationship between treatment and age on bone mineral density, we must control for disease severity. We thus add another (categorical) explanatory variable, disease severity (mild, moderate, and severe). The analysis is now called analysis of covariance (ANCOVA).

- **Repeated-measures ANOVA** is used to assess several paired, or repeated, measurements of the same subjects under different conditions (such as blood pressure measurements taken while the patient is supine, sitting, and standing) or at different points over time (such as muscle strength measured 1, 5, 10, and 20 days after surgery).

Example: Again, building on the previous example, suppose we have measurements of bone mineral density for all patients at the onset of symptoms and at 6 and 12 months after the onset of symptoms. Time can now be added to the ANOVA model as an explanatory variable. Here, time is a repeated measure; although each woman belongs to a single treatment group and to a single age category, each has bone density measurements at three points in time (0, 6, and 12 months).

Error #11. Not Confirming That the Data Met the Assumptions of ANOVA

ANOVA assumes that the response variable is approximately normally distributed within each level of the explanatory variable and that the variability of these distributions is approximately the same. Because most biologic data are not normally distributed,³⁻⁹ the data may need to be mathematically transformed into distributions that are more normally distributed. Alternatively, a nonparametric form of ANOVA can be used. For example, skewed data should probably be analyzed with the Wilcoxon rank-sum test, rather than with one-way ANOVA, and by the Kruskal-Wallis test, rather than with multi-way ANOVA. (The assumptions of

regression analyses are mentioned in Error #9 in the first article of this series.)

Error #12. Not Identifying the Procedure Used to Adjust for Multiple Comparisons in ANOVA

ANOVA is a group comparison that determines whether a statistically significant difference occurs somewhere among the groups studied. If a significant difference occurs, ANOVA is followed by a multiple comparison procedure that compares combinations of groups to determine which groups differ statistically. Common multiple comparison procedures include Tukey's procedure, Student-Neuman-Keuls procedure, Scheffe's method, and Fisher's least-significant method; there are many others.

Error #13. Not Testing the Explanatory Variables for Interaction or Colinearity

Two explanatory variables are said to interact if the effect of one of the response variables depends on the level of the other. For example, alcohol and barbiturates can interact to cause death, even if the amounts of each—by themselves—are not lethal. Interaction implies that the factors should be considered together, not separately. Thus, an analysis of the causes of death from drug overdose would have one factor for blood alcohol level, one for blood barbiturate level, and an interaction term that represents the fact that the effect of alcohol on death depends in part on barbiturate level.

Two variables are said to be colinear if they are highly associated and therefore provide the same information in the model. Systolic and diastolic blood pressure, for example, may contribute such similar information to the model that only one need be used. Testing for interaction and colinearity is usually necessary only in large studies with several explanatory variables.

Error #14. Not Indicating the Goodness-of-Fit of the Model to the Data

Goodness-of-fit indicates how well the model expresses the relationships observed in the data. Examining the residuals (the differences between the observed values and those estimated by the model) helps to determine the fit of the model. The results of the analysis of residuals need not be reported; a statement that the residuals were examined and that the model did (or did not) appropriately fit the data will suffice.

In multiple regression analysis (not ANOVA), the value of R^2 should be reported. This value indicates how much of the variation in the response variable is explained by the factors included in the model. Thus, the higher, the better.

Error #15. Not Reporting Whether and How the Model Was Validated

Multivariate models can be validated or tested against a similar set of data to show that they explain what they seek to explain. One method of validation, used with large samples, is to develop the model on, say, 70% of the data and to compare it with another model based on the remaining 30%. Another method involves removing the data from one subject at a time and recalculating the model. The coefficients and the predictive validity of all the models (there may be hundreds) can then be compared. Such methods are called jackknife procedures. A third method involves developing a new model on a new set of comparable data to determine whether the results are similar.

Errors in Interpreting Differences Between Groups

The majority of biomedical research studies are interested in *differences*, either in one or more groups over time or between two or more groups at the same time. Differences are of interest, for example, when they indicate that one intervention might be more effective than another. Differences can be presented in several forms, however, some of which can be misleading. Here, I describe some of the more common forms, how they can be misinterpreted, and what additional information is needed to prevent these misinterpretations.

Error #16. Not Reporting Confidence Intervals with Estimates

When interpreting any difference, whether it is statistically significant or not, the *direction* and *magnitude* of the difference should be evaluated for its clinical importance. However, because a study is based on a *sample* of the population of interest, rather than on a *census* of the population, its results are actually *estimates* of the differences expected if the study were to be repeated on the entire population. Thus, another factor that should be considered when evaluating differences is the *precision of the estimate*.

In clinical research, the most common measure of precision for an estimate is the 95% confidence interval. In the following example,² evaluating only the estimated size of the difference can be misleading. For this reason, journals now recommend reporting the 95% confidence interval for the difference between groups (that is, for the estimate), instead of, or in addition to, the *P* value for the difference.¹⁰

“*The mean diastolic blood pressure of the treatment group dropped from 110 to 92 mm Hg (P = 0.02).*” This

presentation is the most typical. The pretest and posttest values are given, but not the difference. The mean drop—the 18-mm Hg difference—is statistically significant, but it is also an *estimate* of the drug’s effectiveness, and without a 95% confidence interval, the precision (and therefore the usefulness) of the estimate cannot be determined.

“*The drug lowered diastolic blood pressure by a mean of 18 mm Hg, from 110 to 92 mm Hg (95% CI = 2 to 34 mm Hg; P = 0.02).*” In essence, the confidence interval indicates that if the drug were to be tested on 100 samples similar to the one reported, the average drop in blood pressure would fall between 2 and 34 mm Hg in 95 of the 100 samples. (See *Letter to the Editor and response on page 135.*) A drop of only 2 mm Hg is not clinically important, but a drop of 34 mm Hg is. So, although the *mean* drop in blood pressure in this particular study was statistically significant, the expected difference in blood pressures may not always be clinically important; that is, these study results are actually inconclusive. For conclusive results, more patients probably need to be studied to narrow the confidence interval until *all* or *none* of its values are clinically important.

Error #17. Reporting Only Relative Differences and Not Absolute Ones

The absolute difference between groups is simply the *mathematical difference between their values*, whereas the relative difference is the *absolute difference expressed as a percentage*. By themselves, relative differences can mislead because they can make differences appear to be larger or smaller than they really are.¹¹ For example, a 50% survival rate could mean that two of four patients survived or that 2,000 of 4,000 survived. The absolute difference in survival is two in the smaller study and 2,000 in the larger one. Thus, although both studies show the same relative difference, the absolute difference of the first study is probably too small to justify meaningful conclusions.

In a scientific article, the numerators and denominators should be apparent for all percentages so that the absolute differences can be determined.² This need is especially important when the numbers are less than 100, because the percentages are larger than the actual numbers they represent. “A third of the rats lived, 33% died, and the last one got away.” Here, 33% is one of three rats. In the following, more serious example,¹² readers given the absolute difference usually judge the drug to be far less effective than do readers given the relative difference. “In the Helsinki study of hypercholesterolemic men, after 5 years, 84 of 2030 patients on placebo (4.1%) had heart attacks, whereas only 56 of 2051 men treated

with gemfibrozil (2.7%) had heart attacks ($P < 0.02$)” Here, the **absolute difference** (and therefore, the “absolute risk reduction” in heart attack) was 1.4%; that is, the difference between the frequency of heart attacks in the two groups was 1.4% ($4.1\% - 2.7\% = 1.4\%$). However, the **relative difference** (and therefore, the “relative risk reduction” in heart attack) was 34%; that is, 1.4 is 34% of the 4.1% of men in the control group who had heart attacks ($1.4\%/4.1\% = 34\%$).

Error #18. Not Differentiating Between Unit of Observation and the Number of Patients Improved

The unit of observation or the unit of analysis is what is being studied. In clinical research, the unit of observation is usually the patient. However, sometimes the unit is something other than the patient. The problem comes when, say, differences are reported for the unit of observation but not for the number of patients in whom differences occurred. For example, if a drug markedly improves mean glomerular filtration rate in patients with renal disease, it may also be helpful to know how many patients actually improved.

This issue can be illustrated with a simple example (Figure 1), in which the results can be reported as a mean decrease from time 1 to time 2 or as an increase in two of three (66%!) patients. Both results are technically correct, but reporting only one can be misleading because the mean change is the result of an unusual response in a single patient.

Error #19. Confusing Post-hoc Analyses with Planned Analyses

Post-hoc analyses are analyses performed after investigators have reviewed the study data; that is, post-hoc analyses are *exploratory analyses* suggested by the data and are not planned in advance of data collection. Exploratory analyses are necessary to make the most of the data collection effort. The problem comes when these analyses are presented as planned, primary analyses, rather than as exploratory analyses. Differences detected by post-hoc analyses should be evaluated more critically than differences detected by the planned analyses.

The number of exploratory analyses can sometimes be large. As mentioned in Error #9,¹ generating multiple P values greatly increases the chance of finding a significant P value *somewhere* in the data. Exploratory analyses are thus sometimes called data dredging or “fishing expeditions” when the real search is for *any* significant P value rather than meaningful differences in the data. “Hypothesis-generating studies (sometimes referred to as

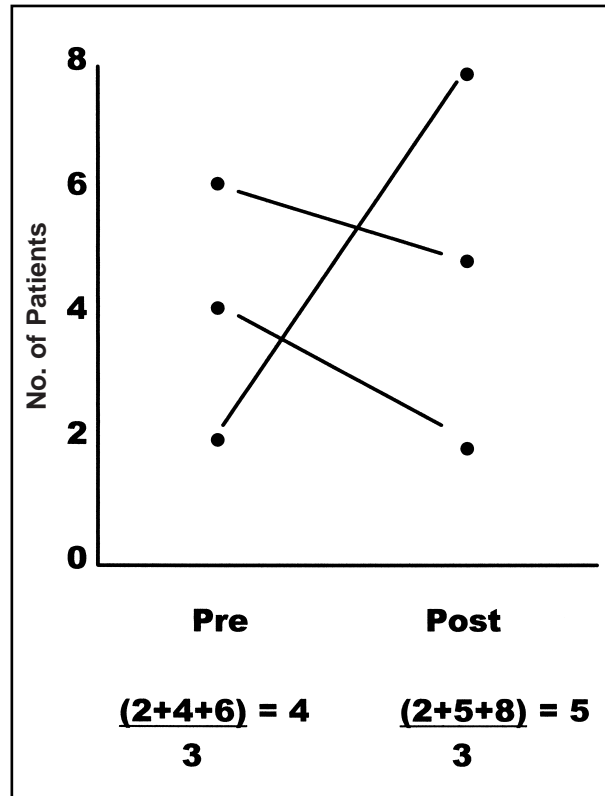


Figure 1

‘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.”¹³

References

1. Common Statistical errors even you can find. Part 1: errors in descriptive statistics and in interpreting probability values. *AMWA J.* 2003;18:67-71.
2. Lang T, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers.* Philadelphia, Pa: American College of Physicians, 1997.
3. Haines SJ. Six statistical suggestions for surgeons. *Neurosurgery.* 1981;9:414-418.
4. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. *Ann Intern Med.* 1981;94(4 part 2):557-592.

5. Evans M, Pollock AV. Trials on trial. A review of trials of antibiotic prophylaxis. *Arch Surg*. 1984;119:109-113.
6. Feinstein AR. X and iprP: an improved summary for scientific communication [editorial]. *J Chron Dis*. 1987;40:283-288.
7. Hall JC, Hill D, Watts JM. Misuse of statistical methods in the Australasian surgical literature. *Aust NZ J Surg*. 1982;52:541-543.
8. Hall JC. The other side of statistical significance: a review of type II errors in the Australian medical literature. *Aust NZ J Med*. 1982;12:7-9.
9. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? *Stat Med*. 1987;6:3-10.
10. Bailar JC, Mosteller F. Guidelines for statistical reporting in articles for medical journals. *Ann Intern Med*. 1988;108:266-273.
11. Guyatt GH, Sackett DL, Cook DJ. Users' guide to the medical literature. II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? *JAMA*. 1994;271:59-63.
12. Brett AS. Treating hypercholesterolemia: how should practicing physicians interpret the published data for patients? *N Engl J Med*. 1989;321:676-680.
13. Mills JL. Data torturing [letter]. *N Engl J Med*. 1993;329:1196-1199.

GUIDEBOOK TO BETTER MEDICAL WRITING

by Robert L. Iles

“The best basic manual on medical writing... everything you need to know about developing a clear, persuasive paper that stands a good chance of publication by a peer-reviewed journal.” Barbara G. Cox, MedEdit Associates, Gainesville, FL. (amazon.com book review)

“Iles has succeeded in boiling down the essentials of medical writing into a cogent handbook.” Linda M. Bonnell, PharmD, *AMWA Journal*, 1999;14:31.

“A concise, no-nonsense approach... provides readers with a series of excellent tips...helpful in my own medical writing and consulting service.” Thomas Buckingham, MD, Bratislava, Slovak Republic. (amazon.com book review)

“Although the focus is on clinical articles, what Iles has to say applies to most scientific writing...” Jude Richard, *CBE Views*, 1999;22:201.

Read an excerpt at www.medwriting.com

Send me _____ copy(ies) at \$ **27.95** ea plus \$3.50 shipping and handling U.S.

25% discount, five or more copies!

Please print

Name _____

Organization _____

Street address _____

City, state, ZIP _____

Enclosed is check money order

Charge to my Visa MasterCard

____ - ____ - ____ - ____

Expiration date _____

Island Press
1065 Wyckford Rd
Olathe, KS 66061
Fax: (913) 782-7138



COMMON STATISTICAL ERRORS EVEN YOU CAN FIND*

PART 3: ERRORS IN DATA DISPLAYS

By Tom Lang, MA

Tom Lang Medical Communications

This article is the third in a series in which I describe several of the more common statistical errors in the biomedical literature. The first article focused on 10 errors in descriptive statistics and in interpreting probability, or P values (*AMWA J.* 2003;18: 67-71), and the second article described 9 errors in interpreting differences between groups (*AMWA J.* 2003;18:103-106). This article addresses 5 errors in presenting statistical information in figures and tables.

Not surprisingly, errors in figures and tables can confuse the interpretation of data. For example, we tend to recall the visual impression of a figure better than the actual message presented by the data. We also tend to compare things that are side-by-side, including data in adjacent columns of a table. As the examples here illustrate, the most effective way (and the only ethical way) to use a figure is to make the visual impression correspond to the message of the data. The best way to use a table is to put the columns to be compared side-by-side.

ERROR #20. Visually distorting relationships on a column chart by starting columns at a baseline value other than zero.

This common error (Figure 1) is sometimes called the *suppressed zero problem*. Unless otherwise informed,

readers assume that the baseline of any chart is zero on the Y axis. To read the chart, readers are supposed to see the values at the tops of the columns.

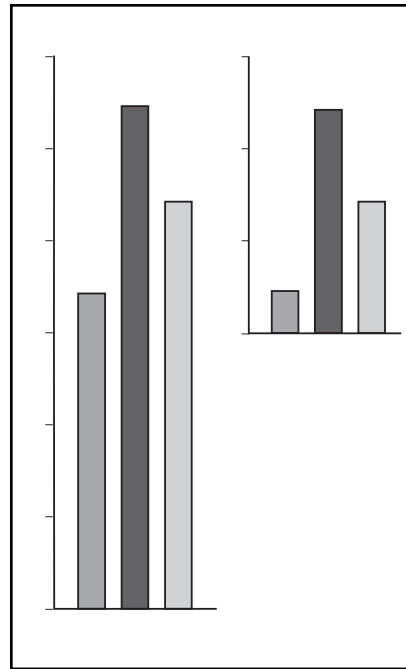


Figure 1. The Suppressed Zero Problem.

The suppressed zero visually distorts the relationships among quantities. Here, A is actually two thirds as large as B, but the suppressed zero makes A appear to be less than one quarter the size of B.

Visually, however, readers actually compare the *heights* of the columns. In the suppressed zero problem, the length of the column is no longer proportional to the value it represents. Thus, in the graph on the right in Figure 1, column A appears to be less than one quarter the size of column B when, in fact, the value of column A is actually two thirds as large as that in column B. To prevent this distortion, the scale and the columns should be “broken” above the expected baseline of zero to indicate clearly that part of the scale has been omitted.

ERROR #21. Visually distorting relationships among data by manipulating the relative scales on the X and Y axes.

Sometimes called the *elastic scale problem*, this error is related to the relationship between the width and height of a graph (Figure 2). By expanding or compressing the scale on the Y axis, differences can be made to look larger or smaller. Expanding or compressing the scale on the X axis can make changes over time to look more sudden or more gradual.

There is no easy way to prevent this

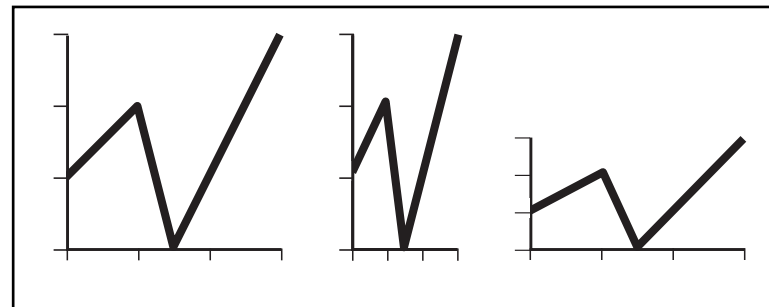


Figure 2. The Elastic Scale Problem. *Uneven scales visually distort relationships among trends. Compressing the scale of the X axis (representing time here) makes changes seem more sudden. Compressing the scale of the Y axis makes the changes seem more gradual. Scales with equal intervals are preferred.*

**This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The AMWA Journal gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.*

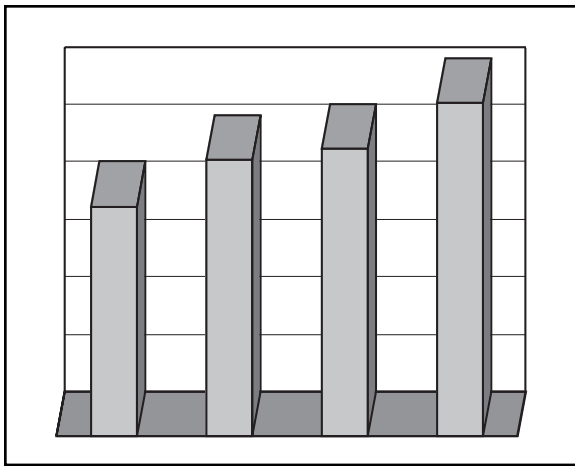


Figure 3. The Double Perspective Problem. *The double perspective problem confuses the reader by shifting the visual reference point, in this case from the back of the column to the front.*

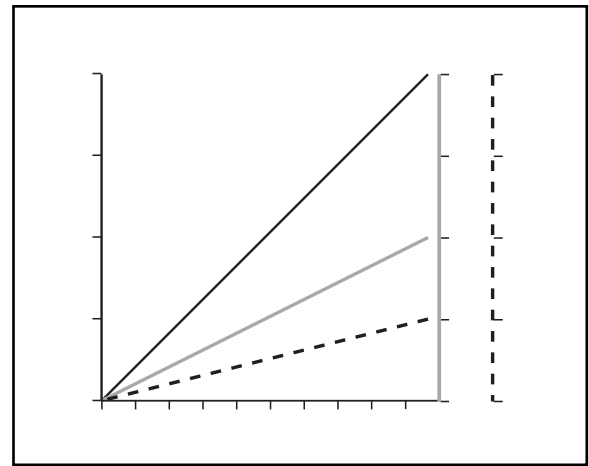


Figure 4. The Double Scale Problem. *Charts with 2 scales, each for a different line of data, can imply a false relationship between the lines, depending on how the scales are presented. Lines A, B, and C represent the same data, but their visual relationships depend on how their respective scales are drawn. Here, Line B seems to increase at half the rate of Line A, whereas Line C seems to increase at one quarter of the rate. Unless the vertical scales are mathematically related, the relationship between the lines can be distorted simply by changing one of the scales.*

error because the scales are likely to be dictated by the data. However, a scale that seems to be unduly compressed or expanded may be a clue that the authors, intentionally or otherwise, are trying to minimize large differences or maximize small differences in the data.

ERROR #22. Visually confusing relationships among data by adding unnecessary dimensions to the chart.

Three-dimensional charts are rarely necessary in biomedical research. The extra dimension, usually added only to make the figure more “attractive,” can mislead readers by directing them to focus on the wrong part of the chart. In Figure 3, it is not clear whether the values should be read from the front or back of the columns. Again, 3-dimensional charts should be examined closely to determine whether the implied visual relationship is actually supported by the data.

ERROR #23. Visually distorting relationships among data by graphing two variables on a single graph using two unrelated vertical scales of measurement.

This *double scale problem* is a problem when the vertical scales can be expanded or compressed *independently* of one another. In Figure 4, lines A, B, and C rise from a value of zero to a value of 4, yet visually, line B appears to be increasing at half the rate of line A and line C appears to be increasing at a quarter of the rate of line A. Thus, the visual comparison between line A and line B can be manipulated by expanding or contracting the vertical scale on which line B is graphed. Graphs with 2 variables graphed on 2 *independent* scales should always be examined closely to determine whether the implied visual relationship is actually supported by the data. (Graphs with related scales, such as gallons on one side and liters on the other have a fixed relationship to one another, so the distortion does not occur.)

ERROR #24. Using tables to store data rather than to communicate information.

The figures and tables used to record data during a study are not necessarily the same ones that should be used to communicate the study’s results. The sample tables show 4 of the 8 possible forms a table might take to compare 3 variables: age, sex, and nationality. All 8 forms of the table would contain the same data. However, the best table for communicating results is the one in which the columns to be compared are placed side-by-side. Thus, Table 1 is preferred for comparing the values of nationality by sex; Table 2 is preferred for comparing the values of men with those of women by nationality; Table 3 of sex by age group; and Table 4, of age groups by sex.

Table 1. Preferred Table Format for Comparing Values in the United States with Those in Japan according to Sex

	Men		Women	
	United States	Japan	United States	Japan
0-21 y				
22-49 y				
≥50 y				

Table 2. Preferred Table Format for Comparing Values of Men with Those of Women according to Nationality

	Japan		United States	
	Men	Women	Men	Women
0-21 y				
22-49 y				
≥50 y				

Table 3. Preferred Table Format for Comparing Values of Men with Those of Women according to Age Group

	0-21 y		22-49 y		≥50 y	
	Men	Women	Men	Women	Men	Women
United States						
Japan						

Table 4. Preferred Table Format for Comparing Values for Age Groups according to Sex

	Men			Women		
	0-21 y	22-49 y	≥50 y	0-21 y	22-49 y	≥50 y
United States						
Japan						

COMMON STATISTICAL ERRORS EVEN YOU CAN FIND!*

PART 4: ERRORS IN CORRELATION AND REGRESSION ANALYSES

By Tom Lang, MA

Tom Lang Medical Communications, Davis, California

This article is the fourth in a series in which I describe several of the more common statistical errors in the biomedical literature. The first article focused on 10 errors in descriptive statistics and in interpreting probability, or P values (*AMWA J.* 2003;18:67-71); the second article described 9 errors in interpreting differences between groups (*AMWA J.* 2003;18:103-106); and the third article addressed 5 errors in presenting statistical information in figures and tables (*AMWA J.* 2004;19:9-11). This article focuses on 3 errors in correlation and regression analyses.

Correlation and regression analyses mathematically describe relationships between continuous variables (variables measured on a continuous scale of equal intervals). In general, 2 variables are considered to be *correlated* when a change in one is likely to be accompanied by a change in the other. The strength of the correlation between variables can be indicated with a correlation coefficient (and its associated confidence interval), which in turn can be tested with hypothesis tests to determine the likelihood that the relationship is the result of chance.¹

Correlation analysis is most easily visualized as a scatter plot in which values from 2 (continuous) variables pertaining to the same subject are plotted on a graph (Figure 1). When

the scatter of data is large and circular, the correlation between the variables is low. The more linear the scatter becomes, the higher the correlation between the variables. A **correlation coefficient** is a number between -1 and +1 that describes the extent of the scatter. A coefficient of 0 means no correlation, whereas a coefficient of -1 means that one value increases linearly as the other value decreases.

Other correlation coefficients include:

- **Pearson's product-moment correlation coefficient, r:** for assessing the relationship between 2 normally distributed, continuous variables
- **Spearman's rank correlation coefficient, rho (ρ):** for assessing the relationship between 2 continuous variables that may or may not be normally distributed
- **Kendall's rank correlation coefficient, tau (τ):** for assessing the relationship between 2 ordinal variables or 1 ordinal and 1 continuous variable
- **Point biserial correlation coefficient:** for assessing the relationship between a continuous variable (age in years) and a categorical variable with 2 levels (such as "recovery status": recovered or not)
- **Point multiserial correlation coefficient:** for assessing the relationship between a continuous variable (total amount of drug administered) and a categorical variable with 3 or more levels (such as "disease severity": mild, moderate, severe)

- **Intraclass and interclass correlation coefficients:** for assessing agreement within and between observers, respectively (often used in assessing diagnostic criteria or patient status)

Correlation analysis for 2 continuous variables can be extended by fitting a "least-squares regression line" to the scatter plot in what is called "simple linear regression analysis." In such an analysis, researchers seek to predict the value of one variable from a known value of the other.

ERROR #25. Confusion when interpreting correlation coefficients

The most common error in interpreting a correlation coefficient is to conclude that changes in 1 variable cause the changes in the other.^{2,3} In fact, correlation is descriptive; it indicates only that 2 characteristics vary together, not that one causes the other to change. Variables that are highly correlated are

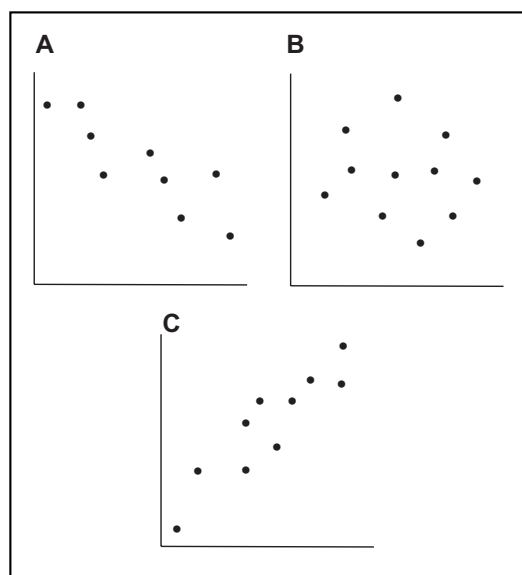


Figure 1. Scatter plots for showing correlation. A. Negative correlation. B. No correlation. C. Positive correlation.

*This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The AMWA Journal gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.

often actually responding to changes in a third variable. For example, in children, shoe size is highly correlated with writing ability. Increasing a child's shoe size does not improve writing ability; rather, both change as a result of age.

Another common error is forgetting that the correlation coefficient must be interpreted.¹ The coefficient has no clinical meaning; it simply describes a relationship in mathematical terms. Correlation is not present or absent, or even low, moderate, or high; any values used to define these categories (such as 0 to 0.3 for low; 0.3 to 0.6 for moderate, and 0.6 to 1 for high) are arbitrary. The coefficient must be interpreted in light of the relationship under study. Thus, a coefficient of 0.9 may be "high" for some relationships and "low" for others, depending on the relationship being studied.

ERROR #26. Not confirming that the relationship in linear regression analysis is linear

As indicated in Error #7—not confirming that the assumptions of statistical tests were met—the assumptions of linear regression analysis need to be confirmed (AMWA J. 2003;18:67-71). The basic assumption of linear regression analysis is that the relationship between 2 variables is linear. This assumption should be assessed mathematically, with an analysis of "residuals." A "residual" is the difference between the observed value of Y and the value of Y predicted by the regression line for a given value of X (Figure 2). When the residuals are small for all values of X and cluster along the 0 point on the Y axis (Figure 3A), the relationship is linear. If the residuals show any other pattern (Figure 3B), the relationship is not linear and the assumptions of linear analysis may have been violated. The graph of residuals need not be shown; a statement that an analysis of residuals confirmed the assumption of linearity is sufficient.¹

ERROR #27. Extending the prediction line beyond the data

When predicting the value of 1 variable, Y, from another, X, the prediction is valid only over the range of X studied; that is, the regression line should not be extended beyond the range of the data. The slope of the line may change greatly at values beyond the studied range (Figure 4). In addition, many relationships have minimum or maximum values, beyond which a regression line could not logically go. For example, a line for predicting a patient's height from weight should not cross either axis, because neither height nor weight can be 0 in this case.¹

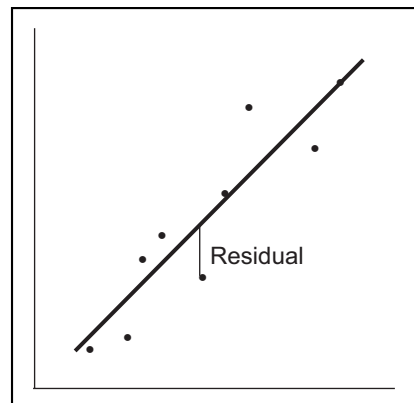


Figure 2. A residual is the difference between the observed value of Y (the dot representing a data point) and the value of Y predicted by the regression line for a given value of X (the Y value where the line crosses the value of X).

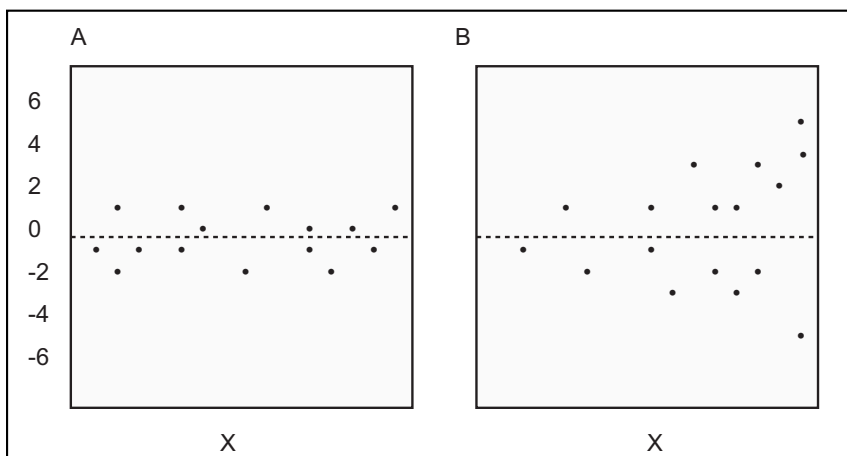


Figure 3. A. When the graphed residuals remain close to 0 over the range of values, the regression line accurately represents the linear relationship of the data. Any other pattern (B) indicates that the relationship may not be linear, which means that linear regression analysis should not be used.

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. Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *BMJ*. 1983;286:1489-1493.
3. Schoolman HM, Becktel JM, Best WR, Johnson AE. Statistics in medical research: principles versus practices. *J Lab Clin Med*. 1968;71(3):357-367.

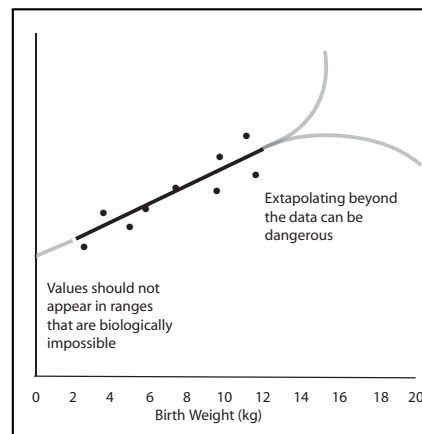


Figure 4. A graph of a simple linear regression analysis showing 2 common errors: extending the regression line beyond the data collected and extending the line into regions where values are not possible.

COMMON STATISTICAL ERRORS EVEN YOU CAN FIND!*

PART 5: ERRORS IN REPORTS OF DIAGNOSTIC TESTS

By Tom Lang, MA

Tom Lang Medical Communications, Davis, California

This article is the fifth in a series in which I describe several of the more common statistical errors in the biomedical literature. The first article focused on 10 errors in descriptive statistics and in interpreting probability, or P values (*AMWA J.* 2003;18:67-71); the second article described 9 errors in interpreting differences between groups (*AMWA J.* 2003;18:103-106); the third article addressed 5 errors in presenting statistical information in figures and tables (*AMWA J.* 2004;19:9-11); and the fourth article focused on 3 errors in correlation and regression (*AMWA J.* 2005;20:10-11). Here, I describe 4 errors in reporting the results of diagnostic tests.

Diagnostic test characteristics—sensitivity, specificity, predictive values, likelihood ratios, and so on—are often misunderstood and are therefore sometimes reported and interpreted incorrectly. In this article, I describe the most common characteristics and how to report them.

ERROR #28. Not defining the meaning or implications of “normal” and “abnormal” test results

A primary purpose of diagnostic testing is to distinguish between “normal” or healthy people and those who have a disease. However, the meaning and implications of “normal” and “abnormal” test results can vary and so they need to be defined. In fact, 6 defini-

tions of “normal” are commonly used in medicine:¹

- A **diagnostic definition** of normal is based on the range of measurements over which the disease is absent and beyond which it is likely to be present. Such a definition is desirable because it is clinically useful. A hematocrit level below 30% is one diagnostic definition of anemia; a level above 50% is one diagnostic definition of polycythemia.
- A **therapeutic definition** of normal is based on the range of measurements over which a therapy is not indicated and beyond which it is beneficial. Again, this definition is clinically useful. Only children of a height below an established threshold might be given human growth hormone to prevent dwarfism, for instance.

Other definitions of normal are less useful, although they are common:

- A **risk factor definition** of normal includes the range of measurements over which the risk of disease is decreased and beyond which the risk is increased. This definition assumes that altering the risk factor alters the actual risk of disease. For example, with rare exceptions, high serum cholesterol is not itself dangerous; only the associated increased risk of heart disease makes a high level “abnormal.”
- A **statistical definition** of normal is based on measurements taken from a disease-free population. This definition usually assumes that the test results are “normally distributed,” that they form a “bell-shaped” curve. The “normal range” is the range of measurements that

includes 2 standard deviations above and below the mean; that is, the range that includes the central 95% of all the measurements. However, the highest 2.5% and the lowest 2.5% of the scores—the “abnormal” scores—have no biologic meaning; they are simply uncommon in a disease-free population. In addition, many biologic test results are not normally distributed, which can make extreme scores more difficult to interpret.

- A **percentile definition** of normal expresses the normal range as the lower (or upper) percentage of the total range. For example, any value in the lower, say, 95% of all observations may be defined as “normal,” and only the upper 5% may be defined as “abnormal.” Again, this definition is based only on the frequency of values and may have no clinical meaning.
- A **social definition** of normal is based on popular beliefs about what is “typical.” Desirable weight or the ability of a child to walk by a certain age, for example, often have social definitions of “normal” that may or may not be medically meaningful.

ERROR #29. Not reporting how uncertain test results were included when calculating the test’s characteristics

Not all diagnostic tests give clear positive or negative results. Perhaps not all of the barium dye was taken; perhaps the bronchoscopy neither ruled out nor confirmed the diagnosis; perhaps observers could not agree on the interpretation of clinical signs. Reporting the number and proportion of non-positive and non-negative results is important because such results affect

**This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The AMWA Journal gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.*

the clinical usefulness of the test.

Uncertain test results may be classified according to 3 types²:

- **Intermediate results** are those that fall between a negative result and a positive result. In a tissue test based on the presence of cells that stain blue, “bluish” cells that are neither unstained nor the required shade of blue might be considered intermediate results.
- **Indeterminate results** are results that indicate neither a positive nor a negative finding. For example, responses on a psychologic test may not determine whether the respondent is or is not alcohol dependent.
- **Uninterpretable results** are produced when a test is not conducted according to specified performance standards. Glucose levels from patients who did not fast overnight may be uninterpretable, for example.

How such results were counted when calculating sensitivity and specificity should be reported. Counting the results as positive or negative or not counting them at all, which often occurs, will cause test characteristics to vary. The standard 2 x 2 table for computing diagnostic sensitivity and specificity does not include rows and columns for uncertain results (Figure 1). Even a highly sensitive or specific test may be of little value if the results are uncertain much of the time.

Test Result	Disease Present	Disease Absent	Total
Positive	a	b	a+b
Negative	c	d	c+d
Total	a+c	b+d	a+b+c+d

Sensitivity = a/a+c
Specificity = d/b+d

If the table reflects the prevalence of disease:
Positive predictive value = a/(a+b)
Negative predictive value = d/(c+d)

Figure 1. Standard table for computing diagnostic test characteristics. The table does not consider uncertain results, which often—and inappropriately—are ignored.

Surprisingly, uncertain results are almost always simply discarded. They are rarely reported, and there is no standard way to incorporate them into the calculations of a test’s characteristics.

Error #30. Confusing sensitivity; specificity; true-positive, false-positive, true-negative, and false-negative results; and positive and negative predictive values

Sensitivity, specificity, and positive and negative predictive values are often misunderstood. The essential difference is that sensitivity and specificity indicate how well the test detects disease *when the patient’s disease status is known*. In contrast, predictive values indicate the *likelihood* that a particular patient will have a disease *if the test result is positive* (the positive predictive value) or will not have the disease *if the test result is negative* (the negative predictive value). Predictive values also assume that the prevalence of disease is known (see later).

Few tests are perfectly sensitive or perfectly specific; most will give false-positive results, false-negative results, or both. Perhaps a more convenient way to remember the true and false, positive and negative combinations is as follows:

True-positive results (sensitivity) indicate *confirmed patients* who now know that they have a disease and can thus be treated appropriately.

True-negative results (specificity) indicate *relieved people* who now know they do not have a disease.

False-positive results indicate *stigmatized people* who will now be treated as having a disease but who are actually healthy.

False-negative results indicate *“stealth” patients* who actually have a disease but who are not believed to have the disease.

ERROR #31. Confusing incidence and prevalence

Incidence is the probability of a new event occurring in a population over a

given period of time. In epidemiologic studies, incidence may be expressed either as a proportion or as a rate.³ Incidence expressed as a proportion is called **cumulative incidence**, which is calculated as:

$$\frac{\text{Number of new cases of disease occurring in a population during a specified period}}{\text{Number of persons in the population at risk for the development of the disease during that period}} \times 1,000$$

For example: *The incidence of the disease was 6002/125,767, or 0.048 per 1,000 people.*

Incidence density is expressed as a rate and uses the concept of “person-years,” or the number of people followed times the number of years that each was followed. Incidence density is calculated as:

$$\frac{\text{Number of new cases in a population}}{\text{Disease-free person-years at risk}} \times 1,000$$

For example: *Among 3 patients, one followed up for 3 years, one for 5 years, and one for 6 years, 1 patient had a relapse. The incidence of relapse is 0.07 (1 case of relapse/14 person-years).*

The key to understanding incidence is to remember that it is the number of new cases in a population that occur during a specified period. Any person in the denominator must have the potential to become part of the numerator.

Prevalence is the number of all the people with a disease (not just new cases) during a period of time divided by the total number of people at risk for the disease during the same time:

$$\text{Prevalence} = \frac{\text{Number of people in the population who have the disease during a specified period}}{\text{Number of people in the population at risk for development of the disease or who have the disease during that period}} \times 1,000$$

References

1. Haynes RB. How to read clinical journals: II. To learn about a diagnostic test. *Can Med Assoc J.* 1981;124:703-10.
2. Simel DL, Feussner JR, Delong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making.* 1987;7:107-14.
3. Gerstman BB. *Epidemiology Kept Simple: An Introduction to Classic and Modern Epidemiology.* New York: Wiley-Liss, 1998.

COMMON STATISTICAL ERRORS EVEN YOU CAN FIND!*

PART 6: ERRORS IN RESEARCH DESIGNS

By Tom Lang, MA

Tom Lang Medical Communications, Davis, California

This article is the sixth in a series in which I describe several of the more common statistical errors in the biomedical literature. The first article focused on 10 errors in descriptive statistics and in interpreting probability, or P values (*AMWA J.* 2003;18:67-71); the second article described 9 errors in interpreting differences between groups (*AMWA J.* 2003;18:103-106); the third article addressed 5 errors in presenting statistical information in figures and tables (*AMWA J.* 2004;19:9-11); the fourth article focused on 3 errors in correlation and regression analyses (*AMWA J.* 2005;20:10-11); and the fifth article addressed 4 errors in reporting the results of diagnostic tests (*AMWA J.* 2005;20:50-51). Here I describe 9 errors in research designs.

To answer a question scientifically, statistical analyses must be combined with an appropriate research design. The research design should specify the question to be answered, how the sample is selected and how large it should be, how patients are assigned to groups, what and how measurements should be taken, and how the data will be analyzed.¹ In addition, the design should prevent bias and confounding factors from interfering with understanding the relationship under study.

*This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The *AMWA Journal* gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.

ERROR #32. Not stating the rationale or purpose for the study

Too often, authors do not explain *why* they conducted the research, assuming instead that it will be obvious to most readers. When the rationale for the study is missing or unclear, the importance of the study cannot be judged.

One common error is what I call LIKA Syndrome (“Little Is Known About . . .”). “Little is known about,” by itself, does not justify a study. For example, “little is known about the relationship between hangnails and schizophrenia (so we had to study it).” Instead, the importance of the problem should be made clear¹:

“Ankylosing spondylitis is an inflammatory arthritis of the spine and joints that affects 350,000 people in the United States. Traditional therapies do not adequately treat these patients, and the need for a safe and effective therapy is substantial.”

Once the importance of the problem has been established, the exact purpose of the research can be stated. However, authors also assume, incorrectly, that readers will know the purpose of the study and so do not state it specifically. Common purposes include:

- Describe a condition or a population
- Identify potential relationships among variables
- Predict the values of one set of variables from another, known set
- Determine the safety of an intervention
- Determine the effects of an intervention under controlled conditions (explanatory trials; see later)
- Determine the effects of an intervention under real-world conditions (pragmatic trials; see later)

- Determine which of two or more interventions is the most effective (studies looking for differences)
- Determine whether one intervention is as good as another (studies looking for similarities; equivalence or noninferiority studies)

ERROR #33. Not distinguishing between “explanatory” and “pragmatic” studies²

Explanatory or **efficacy studies** are performed *to understand* a disease or therapeutic process or to determine how well a treatment works under ideal circumstances. Such studies are best conducted under “optimal” or “laboratory” conditions that allow tight control over patient selection, treatment, and follow-up. The results of such studies may provide insight into underlying biologic mechanisms, but they may not be generalizable to clinical practice, where the setting is not tightly controlled. For example, a double-blind trial of a diagnostic imaging test may be appropriate for evaluating the nature of the test, but because physicians are rarely blinded to related information in clinical practice, the test situation is artificial from a pragmatic standpoint.

Pragmatic or **effectiveness studies**, on the other hand, are performed to *guide decision-making* by determining how well a treatment works in actual application. These studies are usually conducted under “normal” conditions that reflect the circumstances under which medical care is usually provided. The results of such studies may be confounded by any number of factors for which controls were not implemented, which limits their explanatory power but that may enhance their applicability to clinical practice. For

many treatments, outpatient studies are likely to be more realistic than inpatient studies, for example.

One problem with not distinguishing between these two types of studies is that researchers may try to do both in a single study and, as a result, do neither well. Another problem is evaluating one type of study by the standards of the other. For example, a pragmatic definition of a cold might be the presence of at least 3 of 10 cold symptoms, whereas an explanatory definition might require the presence of rhinovirus in the nasal mucus. A pragmatic study might use patient self-reports of efficacy as an outcome, whereas an explanatory study might use trained research nurses to evaluate efficacy.

ERROR #34. Errors in sample selection

Sample selection is among the most important parts of research. If the sample is biased or not representative of the desired population, not only will the results be suspect but the study may also be “fatally flawed,” or so inappropriate that it cannot be corrected by statistical means and is unusable for any purpose.

Sampling is affected by factors that determine

- Which patients are available for study (which may not be representative of the entire population of interest)
- Which patients are actually enrolled in the study (the inclusion-exclusion criteria)

An example of the first group of factors is **referral-filter bias**, which refers to the fact that major medical centers are likely to enroll sicker patients than, say, community hospitals or family physicians. The reason the patients may be sicker is that patients who are not helped by their family physician are referred first to the community hospital and later, if they are still ill, to the major center. Patients treated successfully by their physicians or local hospitals are not

referred and so are “filtered” out of the group available for recruitment.

Inappropriate eligibility criteria are characteristics that can bias or confound the results of a study. A study of the health effects of coffee drinking must account or “control” for the health effects of smoking because many coffee drinkers are also smokers. Thus, comparing a sample of coffee drinkers with a sample of non-coffee drinkers may be biased because the effects of smoking may not be evenly distributed in the two groups. Explanatory studies usually have several inclusion and exclusion criteria so that the patients studied are as alike as possible. In contrast, pragmatic studies usually have fewer criteria because the results are to be applied to a wider range of patients.

ERROR #35. Not considering alternative explanations

Interpreting the results of a study involves identifying probable alternative explanations and then determining which are or are not supported by the evidence. Sometimes, alternative explanations are obvious, such as the need in the above example to separate the health effects of coffee from those of smoking. Other times, they are less obvious, such as a poorly calibrated measuring instrument, a contaminated sample, or an unrecorded protocol violation.

Chance is one common alternative explanation that is often controlled for with the use of P values. Small P values indicate that the evidence for chance as an explanation is weak. However, 5 of every 100 P values will still be significant at the 0.05 level, whether or not the treatment is actually effective.

Another alternative explanation is **time**: patients often get better no matter what is done to them. Would the depression have disappeared without the antidepressants? Would the pain have subsided without the analgesics?

Causal combinations and chains also need to be considered. Sublethal doses of alcohol and barbiturates can

be fatal if taken at the same time; some diseases occur only in the presence of a genetic disposition and a precipitating factor.

One way to rule out alternative explanations is to anticipate them during the study so that they can be “controlled for” statistically. For example, in the US, red cars are involved in accidents significantly more often than cars of any other color. (For this example, it is irrelevant whether the reason is the color itself—maybe red distracts other drivers, maybe people who drive red cars are higher risk takers, or there is some other reason.) What does matter is that although color is known to be associated with differences in accident rates, its effects cannot be determined if color data are not collected in subsequent studies. Age and sex are also associated with many biologic differences and so are routinely collected and analyzed for their effects on results. Not collecting such data prevents analyzing their effects and thus prevents identifying alternative explanations based on age or sex. Finally, biologic plausibility and common sense are sometimes used to rule out alternative explanations.

ERROR #36. Not describing the processes of allocation concealment and group assignment in randomized trials

Random assignment does NOT ensure that the groups will be “equivalent at baseline,” but rather that any differences between them will be the result of chance, rather than systematic error (bias). Bias can affect the assignment process at 3 points. First, the random number sequence may not be truly random. Only sequences generated from a validated random-number computer program or a table of random numbers are suitable. Assignment by odd or even birth dates, hospital admission date, or even coin tossing are not truly random methods. Thus, the method of generating the random number sequence should be reported.¹

Second, once the random number

sequence is associated with patient numbers and group assignment (Table 1), it must be concealed from those enrolling patients into the study. Otherwise, physicians who want their patients to receive the study drug can simply wait to enroll their patient until the next assignment will be to the treatment group.³ To prevent this kind of bias, some studies may use central data coordinating centers to assign patients or include the assignment in sequentially numbered, opaque envelopes (SNOE). When a patient meets the eligibility criteria, the enrolling physician calls the coordinating center or opens the next envelope in the series to determine group assignment.

Table 1. Random Allocation Schedule

Random No.	Patient No.	Group Assignment
023	001	Treatment
592	002	Control
177	003	Treatment
633	004	Treatment
302	005	Control

Allocation concealment is not the same as blinding, the third point at which bias can occur. Blinding refers to keeping group assignment hidden *after* patients are assigned to groups (see later). Blinding is not always possible, but allocation concealment is, and the method used should be reported.¹

ERROR #37. Not reporting who was blinded or the effectiveness of blinding

Blinding is the process of keeping group assignment hidden from the various people involved with the research. Blinding prevents the biases that can occur if treatment status is known: expectation bias on the part of patients and researchers and bias caused by closer monitoring of patients in the treatment group than in the control group, for example.

However, terms like “double blinding” mean different things to different

people.⁴ For this reason who, exactly, is blinded must be reported.^{1,4} Was it patients and caregivers? Patients and data assessors? Patients and statisticians?

Well-conducted studies will also include assessment of the effectiveness of blinding.¹ Patients (and researchers) really do try to determine which group they are in. Maybe the drug has mild side effects; perhaps the placebo has a slightly different taste. Study participants can and do communicate these cues among themselves. If they are successful in “breaking the blind,” they may bias the results.

ERROR #38. Not reporting how data were handled and analyzed

In many trials, especially those with hundreds or thousands of observations, it is important to report how the data were handled. **Data cleaning** is a necessary process of ensuring that all the data are in the correct form in the database. The process may include identifying missing information, generating values to replace missing

information, removing duplicate data, resolving inconsistencies in the data, and detecting and fixing errors. Good practice also dictates that at least a portion of the data be verified for accuracy. **Double data entry** involves entering the data into 2 identical databases and then comparing them for differences. Another method is to check a random sample of, say, 10% of the electronic records against the charts on which data were collected. Reporting any of these methods should increase the credibility of the research.

ERROR #39. Not accounting for all data or all patients or observations¹

Missing data is a common but irritating reporting problem made worse by the thought that the author is careless, lazy, or both. Missing data raise issues about:

- The nature of the missing data. Were extreme values not included in the analysis? Were data lost? Were data ignored because they did not support the hypothesis?
- The generalizability of the present-

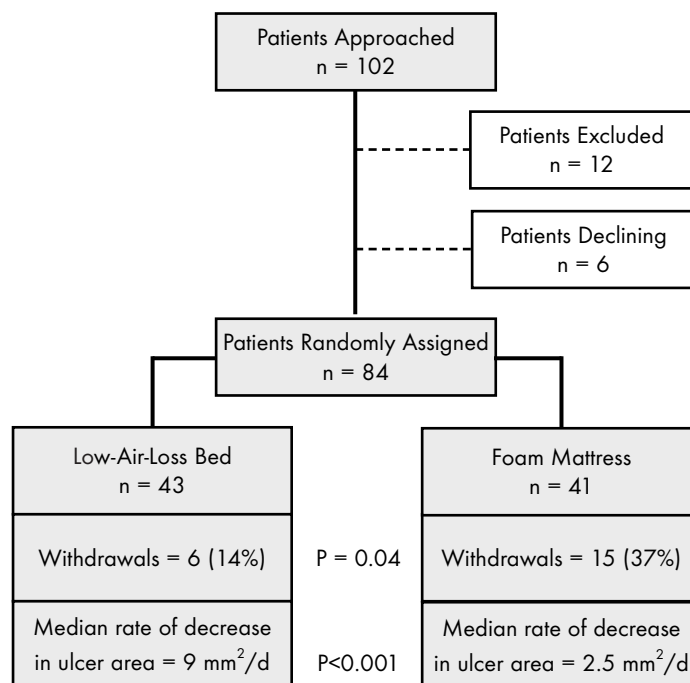


Figure 1. A schematic summary or flow chart of a randomized controlled trial comparing the healing rate of pressure ulcers in patients treated with low-air-loss beds to that of patients treated with foam mattresses.

ed data. Is the range of values reported really the range expected in nature? Is the drop-out rate really that high?

- The quality of the entire study. If the manuscript contains arithmetic errors, how careful was the author about the rest of the research?

One of the most effective ways to account for all patients in research is with a schematic summary or flow chart (Figure 1).¹ Such a visual summary can be used to account for all patients at each stage of the research, to efficiently communicate the study design, and to indicate the denominators for proportions, percentages, and rates. The form of the summary is not as important as the fact that it is visual and a sensible accounting of all patients.

ERROR #40. Not conducting intent-to-treat analysis

Not all patients complete the study as planned. Perhaps they move out of the area, die of other causes, or stop coming for treatment for nonmedical reasons. However, they may also “drop out” of the study because of some aspect of the study itself, such as adverse reactions, dissatisfaction with care, and so on. If these patients are not included in the final analyses, the results will be biased because they do not include the “treatment failures.” Analyses that include only those patients who complete the study are called **on-protocol or per-protocol analysis**. On-protocol analysis is not bad; it is necessary if researchers want to know the effects of actually completing the treatment as planned. However, to protect against the possible bias of patients leaving the study *because of the treatment*, researchers should also perform **intent-to-treat analysis**.³ In intent-to-treat analysis, all patients are analyzed in the groups to which they were assigned, whether or not they completed the study as planned. The 2 analyses can produce different results (Figure 2).

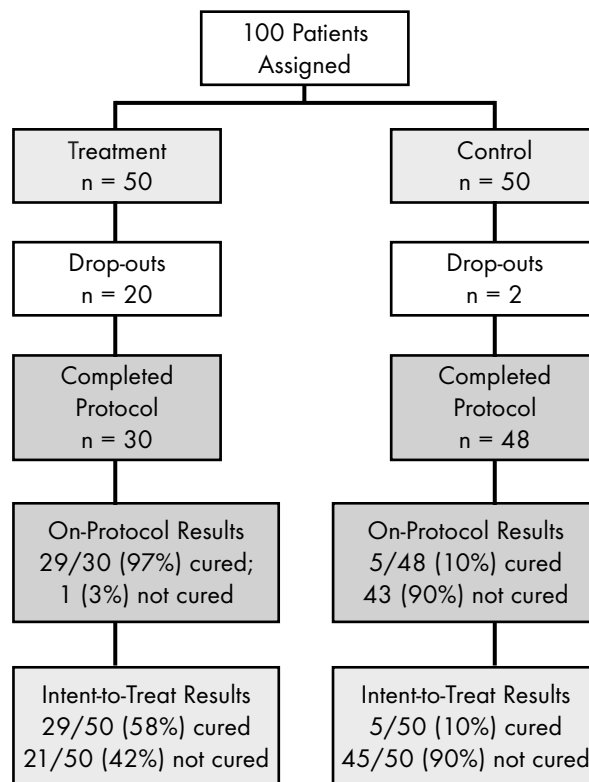


Figure 2. Differences between on-protocol and intent-to-treat analysis. Here, the on-protocol results show a 97% cure rate for patients in the treatment group completing the study as planned. Many patients dropped out of the study, however, perhaps because of the treatment, so the intent-to-treat analysis shows a cure rate of only 58%, although it is still better than the rate for the control group.

References

1. Moher D, Schulz KF, Altman DG, for the CONSORT Group. The CONSORT Statement: revised recommendations for improving the results of parallel-design randomized trials. *Ann Intern Med.* 2001;134:657-662.
2. Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis.* 1967;20:637-648.
3. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA.* 1995;273:408-412.
4. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. *Lancet.* 2002;359:696-700.

COMMON STATISTICAL ERRORS EVEN YOU CAN FIND!*

PART 7: COMMON ERRORS IN CONCLUSIONS

By Tom Lang, MA

Tom Lang Medical Communications, Davis, CA

This article is the last in a series in which I describe several of the more common statistical errors in the biomedical literature. The first article focused on 10 errors in descriptive statistics and in interpreting probability, or P values (*AMWA J.* 2003;18: 67-71); the second article described nine errors in interpreting differences between groups (*AMWA J.* 2003;18: 103-106); the third addressed five errors in presenting statistical information in figures and tables (*AMWA J.* 2004;19:9-11); the fourth focused on three errors in correlation and regression analyses (*AMWA J.* 2005;20:10-11); the fifth addressed four errors in reporting the results of diagnostic tests (*AMWA J.* 2005;20:50-51); and the sixth addressed nine errors in research designs (*AMWA J.* 2005;20:112-115). Here I describe four errors that occur in reports of results and conclusions.

ERROR #41. Important limitations of the study are not acknowledged or explained

All studies have limitations. Unacknowledged limitations do science no good and can cast suspicion on the entire study if they are discovered after publication. In addition, acknowledged limitations usually add credibility to the study (in the absence of fatal flaws). For this reason, most research articles should include in the Discussion the subheading, "Limitations of

**This series is based on 10 articles first translated and published in Japanese by Yamada Medical Information, Inc. (YMI, Inc.), of Tokyo, Japan. Copyright for the Japanese articles is held by YMI, Inc. The AMWA Journal gratefully acknowledges the role of YMI in making these articles available to English-speaking audiences.*

the Study." Conclusions must be drawn in light of the study's limitations. Common limitations of studies include the following.

- Desired sample size was not obtained
- Sample was not representative
- Study did not go as planned
- Drop-out rate was higher than acceptable
- Number of events was inadequate for statistical purposes
- Bias or confounding (alternative explanations) cannot be ruled out
- Outlying or missing data may have affected the analysis
- Reliability or validity of measurements was lower than expected
- Baseline risk (the event rate in the control group) was too low to show an effect (*see later*)

ERROR #42. The conclusions are inconsistent with the research

Conducting, interpreting, and communicating research are difficult tasks. The complexity of the study, the opinions of multiple authors, conflicting comments from reviewers, long delays during portions of the research, and shifting priorities among authors can result in inconsistencies in the published article. For example, in a surprising number of articles, the conclusions reported in the Abstract are different from those reported in the Discussion. For this reason, it is important to determine whether the conclusions and interpretations of the study are consistent with the

- Original study question
- Research as planned
- Research as conducted
- Results as reported

Scientific writing is, by definition, persuasive writing. But the persuasion

must be based on fact and logic, not supposition or speculation. Speculation is necessary and appropriate in many articles, but only when labeled as such. Each conclusion should be supported in the article by facts established in the study itself or in other, cited, research and by logical, biologically plausible reasoning.

ERROR #43. Not accounting for baseline risk when forming conclusions

Baseline risk is the risk of the illness occurring in the comparison population. In many clinical trials, this baseline risk is represented by how often it occurs in the control group (the "event rate" in the control group). For a treatment to be effective, it must reduce the risk of the illness occurring in the treatment group below that in the control group.

It is difficult to prevent what isn't happening, however: to be able to prevent death, people have to be dying. So, if the event rate in the control group is small, the treatment effect is also likely to be small, not necessarily because the treatment is ineffective but because it did not have the opportunity to be effective in this particular group of patients.¹ Alternatively, when the event rate in the control group is high, the opportunity to be effective is also greater, and effect sizes may be correspondingly higher, even for the same treatment. Thus, researchers need to consider the baseline risk when interpreting the results of their own study, as well as those of similar studies of the same treatment.

ERROR #44. Forgetting that the form in which results are reported influences their interpretation

The results of medical studies are often

mathematical expressions, such as correlation coefficients, P values, odds ratios, or measures of risk. Unless readers are aware of how, exactly, these mathematical expressions should be interpreted, their understanding of the results can be influenced by the form in which the results are reported.

Reporting the results in clinically interpretable terms, such as “effort-to-yield” measures, helps to avoid this problem.

Effort-to-yield measures indicate how much clinical “effort” is needed to produce one more unit of benefit or to create one more unit of harm. The most common measures are the number needed to treat (NNT) and the number needed to harm (NNH). For example, in a study of anticoagulants, the NNT is the number of patients who must be treated to prevent one additional heart attack, whereas the NNH is the number who can be treated before one serious bleeding event occurs.

Results of the Helsinki study of hypercholesterolemic men (see Error #17 [AMWA J. 2003;18:103-106]) provide an excellent example of how the form in which results are reported can influence interpretation.²

- **Results expressed as the absolute risk reduction**

In the Helsinki study of hypercholesterolemic men, after 5 years, the absolute risk reduction in the gemfibrozil-treated group was 1.4%.

- **Results expressed as the relative risk reduction**

In the Helsinki study of hypercholesterolemic men, after 5 years, the relative risk reduction in the gemfibrozil-treated group was 34%.

- **Results expressed as the number needed to treat**

The results of the Helsinki study of 4,081 hypercholesterolemic men indicate that 71 men would need to be treated for 5 years to prevent a single heart attack.

- **Results expressed in another effort-to-yield measure**

In the Helsinki study of 4,081 hypercholesterolemic men, after 5 years, the results indicate that about 200,000 doses of gemfibrozil were taken for each heart attack prevented.

Other examples of interpretation errors have been presented throughout this series of articles.

- Interpreting nonstatistically significant results as “negative” when they are, in fact, inconclusive (Error #8 [AMWA J. 2003;18:67-71])
- Confusing statistical significance, as indicated by P values, with biologic importance (Error #10 [AMWA J. 2003;18:67-71])
- Interpreting differences between groups as clinically important, without knowing whether the 95% confidence interval contains only clinically important values (Error #16 [AMWA J. 2003;18:103-106])
- Interpreting large percent differences between groups as important when the absolute differences are small (Error #17 [AMWA J. 2003;18:103-106])
- Assuming that correlation is causation (Error #25 [AMWA J. 2005;20:10-11])
- Assuming that a positive diagnostic test result indicates a medical abnormality (Error #28 [AMWA J. 2005;20:10-11])
- Interpreting the results of a “pragmatic study” by the criteria of “explanatory” studies, and vice versa (Error #33 [AMWA J. 2005;20:112-115])
- Accepting an obvious or desired explanation of results when alternative explanations have not been ruled out (Error #35 [AMWA J. 2005;20:112-115])

CONCLUSION

The problem of poor statistical reporting is long-standing, widespread, potentially serious, and almost unknown, despite the fact that most errors concern basic statistical concepts and can be easily avoided by following a few guidelines.^{3,4}

Good analytical thinking, supplemented by a knowledge of research design and statistical analyses, a healthy skepticism of all research, and the ability to ignore the reputations of the authors, their institutions, and the publishing journal are all necessary to identify problems in biomedical research. The 44 errors described in this series of articles are certainly among the more common errors made in reporting clinical research, but there are many more. Be careful!

References

1. Ioannidis JB, Lau J. The impact of high-risk patients on the results of clinical trials. *J Clin Epidemiol.* 1997;50:1089-1098.
2. Brett AS. Treating hypercholesterolemia: how should practicing physicians interpret the published data for patients? *N Engl J Med.* 1989;321:676-680.
3. Gore SM, Jones G, Thompson SG. The Lancet's statistical review process: areas for improvement by authors. *Lancet.* 1992;340:100-102.
4. Lang T, Secic M. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers.* Philadelphia, PA: American College of Physicians;1997.