

Twenty Statistical Errors Even *YOU* Can Find in Biomedical Research Articles

Tom Lang

Tom Lang Communications, Murphys, Ca, USA

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

"Good research deserves to be presented well, and good presentation is as much a part of the research as the collection and analysis of the data. We recognize good writing when we see it; let us also recognize that science has the right to be written well." (2)

Statistical probability was first discussed in the medical literature in the 1930s (3). 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 (4-7). The problem of poor statistical reporting is, in fact, longstanding, widespread, potentially serious, and not well known, despite the fact that most errors concern basic statistical concepts and can be easily avoided by following a few guidelines (8).

The problem of poor statistical reporting has received more attention with the growth of the evidence-based medicine movement. Evidence-based medicine is literature-based medicine and depends on the quality of published research. As a result, several groups have proposed reporting guidelines for different types of trials (9-11), and a comprehensive set of guidelines for reporting statistics in medicine has been compiled from an extensive review of the literature (12).

Here, I describe 20 common statistical reporting guidelines that can be followed by authors, editors, and reviewers who know little about statistical analysis. These guidelines are but the tip of the iceberg: readers wanting to know more about the iceberg should consult more detailed texts (12), as well as the references cited here. To keep the tension mounting in an often dull subject, the guidelines are presented in order of increasing importance.

The guidelines described here are taken from *How To Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers*, by Thomas A. Lang and Michelle Secic (American College of Physicians, 1997).

Error #1: Reporting measurements with unnecessary precision

Most of us understand numbers with one or two significant digits more quickly and easily than numbers with three or more digits. Thus, rounding numbers to two significant digits improves communication (13). For instance, in the sentences below, the final population size is about three times the initial population size for both the women and the men, but this fact is only apparent after rounding:

– The number of women rose from 29,942 to 94,347 and the number of men rose from 13,410 to 36,051.

– The number of women rose from 29,900 to 94,300 and the number of men rose from 13,400 to 36,000.

– The number of women rose from about 30,000 to 94,000 and the number of men rose from about 13,000 to 36,000.

Many numbers do not need to be reported with full precision. If a patient weighs 60 kg, reporting the weight as 60.18 kg adds only confusion, even if the measurement was that precise. For the same reason, the smallest *P* value that need be reported is $P < 0.001$.

Error #2: Dividing continuous data into ordinal categories without explaining why or how

To simplify statistical analyses, continuous data, such as height measured in centimeters, are often separated into two or more ordinal categories, such as short, normal, 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 (12). In some cases, the boundaries (or "cut points")

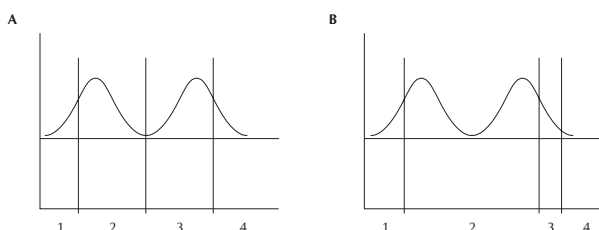


Figure 1. Authors should state why and how continuous data were separated into ordinal categories to avoid possible bias. **A.** For this distribution, these categories appear to have been reasonably created. **B.** The rationale for creating these categories should be explained.

that define the categories can be chosen to favor certain results (Fig. 1).

Error #3: Reporting group means for paired data without reporting within-pair changes

Data taken from the same patient are said to be "paired." In a group of patients with data recorded at two time points, differences can occur between the group means over time, as well between each individual's measurements over time. However, changes in the individuals' measurements may be hidden by reporting only the group means (Fig. 2). Unless the individual data are reported, readers may not know about conflicts between the two measures. The results in Figure 2, for example, can be reported as a mean decrease from time 1 to time 2 or as an increase in two of three patients. Both results are technically correct, but reporting only one can be misleading.

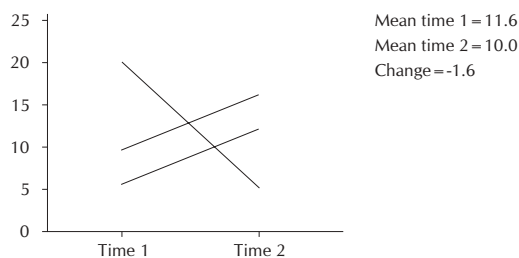


Figure 2. Paired data should be reported together so that the changes within each patient, as well as in group means, can be evaluated. Here, the results can be reported as a mean drop of 1.6 units or that units increased in 2 of 3 patients.

Error #4: Using descriptive statistics incorrectly

Two of the most common descriptive statistics for continuous data are the *mean* and *standard deviation*. However, these statistics correctly describe only a "normal" or "Gaussian" distribution of values. By definition, about 68% of the values of a normal distribution are within plus or minus 1 standard deviation of the mean, about 95% are within plus or minus 2 standard deviations, and about 99% are within plus or minus 3 standard deviations. In markedly non-normal distributions, these relationships are no longer true, so the mean and standard deviation do not communi-

cate the shape of the distribution well. Instead, other measures, such as the *median* (the 50th percentile: the value dividing the data into an upper and a lower half) and *range* (usually reported by giving the minimum and maximum values) or *interquartile range* (usually reported by giving the 25th and the 75th percentiles) are recommended (14).

Although the mean and standard deviation can be calculated from as few as two data points, these statistics may not describe small samples well. In addition, most biological data are not normally distributed (15). For these reasons, the median and range or interquartile range should probably be far more common in the medical literature than the mean and standard deviation.

Error #5: Using the standard error of the mean (SEM) as a descriptive statistic or as a measure of precision for an estimate

The mean and standard deviation describe the center and variability of a normal distribution of a characteristic for a *sample*. The *mean* and *standard error of the mean (SEM)*, however, are an estimate (the mean) and a measure of its precision (the SEM) for a characteristic of a *population*. However, the SEM is always smaller than the standard deviation, so it is sometimes reported instead of the standard deviation to make the measurements look more precise (16). Although the SEM is a measure of precision for an estimate (1 SEM on either side of the mean is essentially a 68% confidence interval), the preferred measure of precision in medicine is the *95% confidence interval* (17). Thus, the mean and SEM can sometimes refer to a sample and sometimes to a population. To avoid confusion, the mean and standard deviation are the preferred summary statistics for (normally distributed) data, and the mean and 95% confidence interval are preferred for reporting an estimate and its measure of precision.

For example, if the mean weight of a sample of 100 men is 72 kg and the SD is 8 kg, then (assuming a normal distribution), about two-thirds of the men (68%) are expected to weigh between 64 kg and 80 kg. Here, the mean and SD are used correctly to describe this distribution of weights.

However, the mean weight of the sample, 72 kg, is also the best estimate of the mean weight of all men in the population from which the sample was drawn. Using the formula $SEM = SD/\sqrt{n}$, where $SD = 8$ kg and $n = 100$, the SEM is calculated to be 0.8. The interpretation here is that if similar (random) samples were repeatedly drawn from the same population of men, about 68% of these samples would be expected to have mean values between 71.2 kg and 72.8 kg (the range of values between 1 SEM above and below the estimated mean).

The preferred expression for an *estimate and its precision* is the mean and the 95% confidence interval (the range of values about 2 SEMs above and below the mean). In the example here, the expression would be "The mean value was 72 mg (95% CI = 70.4 to 73.6 mg)," meaning that if similar (random)

samples were repeatedly drawn from the same population of men, about 95% of these samples would be expected to have mean values between 70.4 mg and 73.6 mg.

Error #6: Reporting only P values for results

P values are often misinterpreted (18). Even when interpreted correctly, however, they have some limitations. For main results, report the absolute difference between groups (relative or percent differences can be misleading) and the 95% confidence interval for the difference, instead of, or in addition to, *P* values. The sentences below go from poor to good reporting:

– “The effect of the drug was statistically significant.” This sentence does not indicate the size of the effect, whether the effect is clinically important, or how statistically significant the effect is. Some readers would interpret “statistically significant” in this case to mean that the study supports the use of the drug.

– “The effect of the drug on lowering diastolic blood pressure was statistically significant ($P < 0.05$)” Here, the size of the drop is not given, so its clinical importance is not known. Also, *P* could be 0.049; statistically significant (at the 0.05 level) but so close to 0.05 that it should probably be interpreted similarly to a *P* value of, say, 0.51, which is not statistically significant. The use of an arbitrary cut point, such as 0.05, to distinguish between “significant” and “non significant” results is one of the problems of interpreting *P* values.

– “The mean diastolic blood pressure of the treatment group dropped from 110 to 92 mm Hg ($P = 0.02$).” This sentence is perhaps the most typical. The pre- 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, 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 in 95 of those 100 samples would probably range between 2 and 34 mm Hg. 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 pressures in this study was statistically significant, the expected difference in blood pressures in other studies may not always be clinically important; that is, the study is inconclusive.

When a study produces a confidence interval in which *all* the values are clinically important, the intervention is much more likely to be clinically effective. If *none* of the values in the interval are clinically important, the intervention is likely to be ineffective. If only some of the values are clinically important, the study probably did not enroll enough patients.

Error #7: Not confirming that the data met the assumptions of the statistical tests used to analyze them

There are hundreds of statistical tests, and several may be appropriate for a given analysis. However, tests may not give accurate results if their assumptions are not met (19). For this reason, both the name of the test and a statement that its assumptions were met should be included in reporting every statistical analysis. For example: “The data were approximately normally distributed and thus did not violate the assumptions of the *t* test.”

The most common problems are:

– Using parametric tests when the data are not normally distributed (skewed). In particular, when comparing two groups, Student’s *t* test is often used when the Wilcoxon rank-sum test (or another non-parametric test) is more appropriate.

– Using tests for independent samples on paired samples, which require tests for paired data. Again, Student’s *t* test is often used when a paired *t* test is required.

Error #8: Using linear regression analysis without establishing that the relationship is, in fact, linear

As stated in Guideline #7, every scientific article that includes a statistical analysis should contain a sentence confirming that the assumptions on which the analysis is based were met (12). This confirmation is especially important in linear regression analysis, which assumes that the relationship between a response and an explanatory variable is linear. If this assumption is not met, the results of the analysis may be incorrect.

The assumption of linearity may be tested by graphing the “residuals”: the difference between each data point and the regression line (Fig. 3). If this graph is flat and close to zero (Fig. 4A), the relationship is linear. If the graph shows any other pattern, the relationship is not linear (Fig. 4B, 4C, and 4D.) Testing the assumption of linearity is important because simply looking at graphed data can be misleading (Fig. 5).

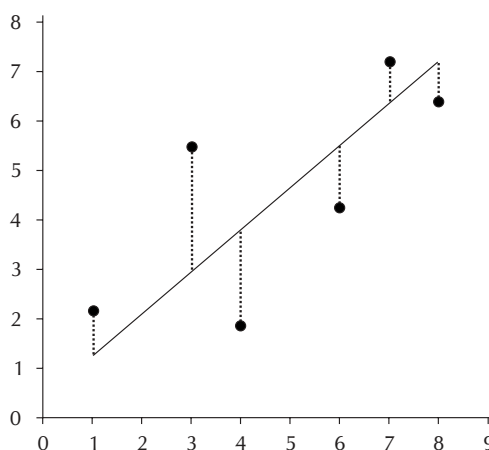


Figure 3. A residual is the distance between an actual, observed value and the value predicted by the regression line.

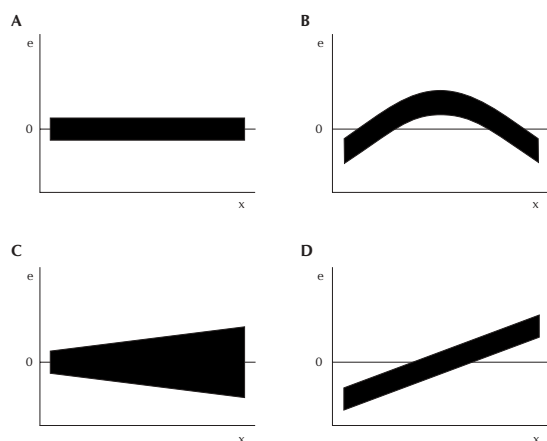


Figure 4. A. When the graphed residuals remain close to zero over the range of values, the regression line accurately represents the linear relationship of the data. Any other pattern (B, C, and D) indicates that the relationship is not linear, which means that linear regression analysis should not be used.

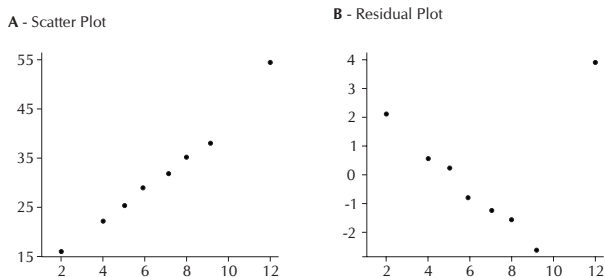


Figure 5. The appearance of linearity in a set of data can be deceptive. Here, a relationship that appears to be linear (A) is obviously not, as indicated by the graph of residuals (B).

Error #9: Not accounting for all data and all patients

Missing data is a common but irritating reporting problem made worse by the thought that the author is careless, lazy, or both (20). Missing data raise issues about:

- the nature of the missing data. Were extreme values not included in the analysis? Were data lost in a lab accident? Were data ignored because they did not support the hypothesis?
- the generalizability of the presented data. Is the range of values really the range? Is the drop-out rate really that low?
- the quality of entire study. If the totals don't match in the published article, how careful was the author during the rest of the research?

One of the most effective ways to account for all patients in a clinical trial is a flow chart or schematic summary (Fig. 6) (9,12,21). Such a visual summary can account for all patients at each stage of the trial, efficiently summarize the study design, and indicate the probable denominators for proportions, percentages, and rates. Such a graphic is recommended by the CONSORT Statement for reporting randomized trials (9).

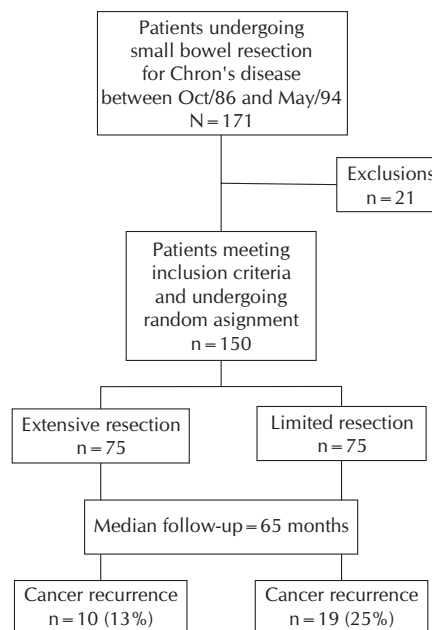


Figure 6. A flow chart of a randomized clinical trial with two treatment arms, showing the disposition of all patients at each stage of the study.

Error #10: Not reporting whether or how adjustments were made for multiple hypothesis tests

Most studies report several *P* values, which increases the risk of making a type I error: such as saying that a treatment is effective when chance is a more likely explanation for the results (22). For example, comparing each of six groups to all the others requires 15 "pair-wise" statistical tests – 15 *P* values. 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 (12):

- *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 endpoint measured at several different times).
- *comparing groups at multiple time points* with a series of individual group comparisons.

Multiple testing is often desirable, and exploratory analyses should be reported as exploratory. "Data dredging," however – undisclosed analyses involving computing many *P* values to find *something* that is statistically significant (and therefore worth reporting) – is considered to be poor research.

Error #11: Unnecessarily reporting baseline statistical comparisons in randomized trials

In a true randomized trial, each patient has a known and usually equal probability of being assigned to either the treatment or the control group. Thus, any differences between groups at baseline are, by definition, the result of chance. Therefore, significant differences in baseline data (Table 1) do not indicate bias (as they might in other research designs) (9). Such comparisons may indicate statistical imbalances between the groups that may need to be taken into account later in the analysis, but the *P* values do not need to be reported (9).

Table 1. Statistical baseline comparisons in a randomized trial. By chance, the groups differ in median albumin scores ($P=0.03$); the difference does not indicate selection bias. Here, *P* values need not be reported for this reason

Variable	Control (n=43)	Treatment (n=51)	Difference	<i>P</i>
Median age (years)	85	84	1	0.88
Men (n, %)	21 (49)	21 (51)	3%	0.99
Median albumin (g/L)	30.0	33.0	3.0 g/L	0.03
Diabetes (n,%)	11 (26)	8 (20)	6%	0.83

Assuming that alpha is set at 0.05, of every 100 baseline comparisons in randomized trials, 5 should be statistically significant, just by chance. However, one study found that among 1,076 baseline comparisons in 125 trials, only 2% were significant at the 0.05 level (23).

Error #12: Not defining "normal" or "abnormal" when reporting diagnostic test results

The importance of either a positive or a negative diagnostic test result depends on how "normal" and "abnormal" are defined. In fact, "normal" has at least six definitions in medicine (24):

- 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 of normal is desirable because it is clinically useful.

- 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.

Other definitions of normal are perhaps less useful for patient care, although they are unfortunately common:

- A *risk factor definition* of normal includes the range of measurements over which the risk of disease is not increased 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 two 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 clinical meaning; they are simply uncommon. Unfortunately, many test results are not normally distributed.

- 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 test results may be defined as "normal," and only the upper 5% may be defined as "abnormal." Again, this definition is based on the frequency of values and may have no clinical meaning.

- A *social definition* of normal is based on popular beliefs about what is normal. 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 important.

Error #13: Not explaining how uncertain (equivocal) diagnostic test results were treated when calculating the test's characteristics (such as sensitivity and specificity)

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 the clinical usefulness of the test.

Uncertain test results may be one of three types (25):

- *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 psychological 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. Test characteristics will vary, depending on whether the

results are counted as positive or negative or were not counted at all, which is often the case. The standard 2x2 table for computing diagnostic sensitivity and specificity does not include rows and columns for uncertain results (Table 2). Even a highly sensitive or specific test may be of little value if the results are uncertain much of the time.

Table 2. Standard table for computing diagnostic test characteristics*

Test result	Disease		Totals
	present	absent	
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. Likelihood ratios can also be calculated from the table. The table does not consider uncertain results, which often – and inappropriately – are ignored.

Error #14: Using figures and tables only to “store” data, rather than to assist readers

Tables and figures have great value in storing, analyzing, and interpreting data. In scientific presentations, however, they should be used to *communicate information*, not simply to “store” data (26). As a result, published tables and figures may differ from those created to record data or to analyze the results. For example, a table presenting data for 3 variables may take any of 8 forms (Table 3). Because numbers are most easily compared side-by-side, the most appropriate form in Table 3 is the one in which the variables to be compared are side-by-side. That is, by putting the variables to be compared side-by-side, we encourage readers to make a specific comparison.

The table and images in Figure 7 show the same data: the prevalence of a disease in nine areas. How-

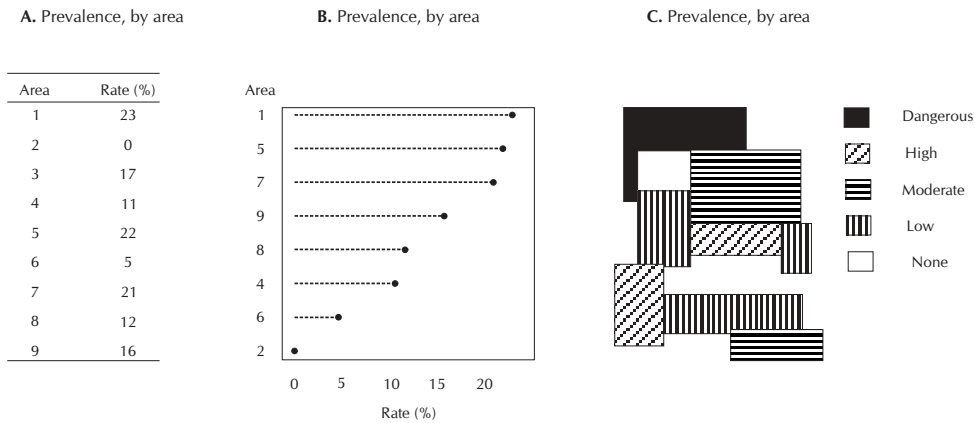


Figure 7. Tables and figures should be used to communicate information, not simply to store data. **A.** Tables are best for communicating or referencing precise numerical data. **B.** Dot charts are best for communicating general patterns and comparisons. **C.** Maps are best for communicating spatial relationships.

Table 3. A table for reporting 3 variables (nationality, sex, and age group) may take any of 8 forms:

Form 1

Age (years)	Men		Women	
	US	China	US	China
0-21				
22-49				
50+				

Form 2

Age (years)	China		US	
	men	women	men	women
0-21				
22-49				
50+				

Form 3

	0-21 years		22-49 years		50+ years	
	men	women	men	women	men	women
US						
China						

Form 4

	Men (age, years)			Women (age, years)		
	0-21	22-49	50+	0-21	22-49	50+
US						
China						

Form 5

	0-21 years		22-49 years		50+ years	
	US	China	US	China	US	China
Men						
Women						

Form 6

	US (age, years)			China (age, years)		
	0-21	22-49	50+	0-21	22-49	50+
Men						
Women						

Form 7

	0-21 years	22-49 years	50+ years
Men:			
US			
China			
Women:			
US			
China			

Form 8

	0-21 years	22-49 years	50+ years
US:			
men			
women			
China:			
men			
women			

ever, the table is best used to communicate and to reference precise data; the dot chart, to communicate how the areas compare with one another; and the map, to communicate the spatial relationships between the areas and disease prevalence.

Error #15: Using a chart or graph in which the visual message does not support the message of the data on which it is based

We remember the visual message of an image more than the message of the data on which it is based (27). For this reason, the image should be adjusted until its message is the same as that of the data. In the “lost zero” problem (Fig. 8A), column 1 appears to be less than half as long as column 2. However, the chart is misleading because the columns do not start at zero: the zero has been “lost.” The more accurate chart, showing the baseline value of zero (Fig. 8B), shows that column 1 is really two-thirds as long as column 2. To prevent this error, the Y axis should be “broken” to indicate that the columns do not start at zero (Fig. 8C).

In the “elastic scales” problem, one of the axes is compressed or lengthened disproportionately with respect to the other, which can distort the relationship between the two axes (Fig. 9). Similarly, in the “double scale” problem, unless the scale on the right has

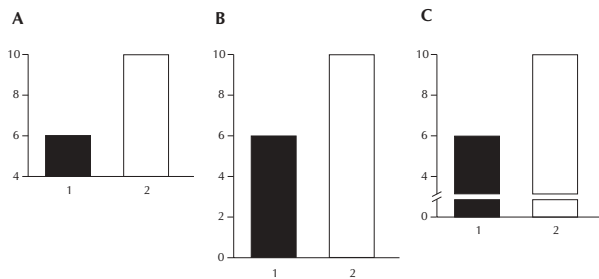


Figure 8. A. Charts and graphs that do not begin at zero can create misleading visual comparisons. B. Here, the actual length of both columns can be compared accurately. C. When space prohibits starting with zero as a baseline, the axis should be “broken” to indicate that the baseline is not zero.

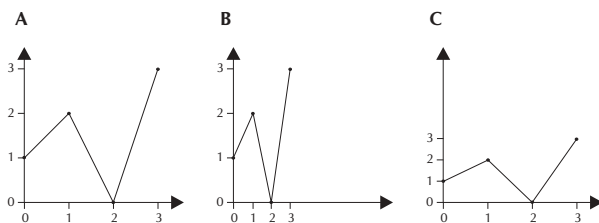


Figure 9. Uneven scales can visually distort relationships among trends. Compressing the scale of the X axis (representing time in this example) makes changes seem more sudden. Compressing the scale of the Y axis makes the changes seem more gradual. Scales with equal intervals are preferred.

some mathematical relationship with the scale on the left, the relationship between two lines can be distorted (Fig. 10).

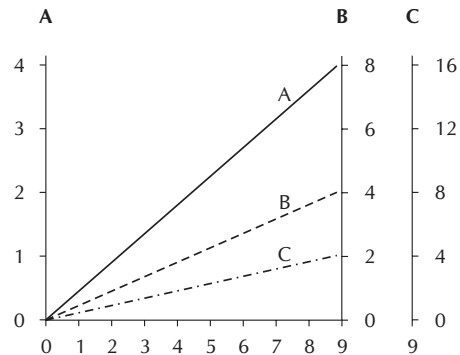


Figure 10. Charts with two scales, each for a different line of data, can imply a false relationship between the lines, depending on how the scales are presented. Lines A, B, and C represent the same data, but their visual relationships depend on how their respective scales are drawn. Here, Line B seems to increase at half the rate of Line A, whereas Line C seems to increase at a 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 #16: Confusing the “units of observation” when reporting and interpreting results

The “unit of observation” is what is actually being studied. Problems occur when the unit is something other than the patient. For example, in a study of 50 eyes, how many patients are involved? What does a 50% success rate mean?

If the unit of observation is the heart attack, a study of 18 heart attacks among 1,000 people has a sample size of 18, not 1,000. The fact that 18 of 1,000 people had heart attacks may be important, but there are still only 18 heart attacks to study.

If the outcome of a diagnostic test is a judgment, a study of the test might require testing a sample of judges, not simply a sample of test results to be judged. If so, the number of judges involved would constitute the sample size, rather than the number of test results to be judged.

Error #17: Interpreting studies with nonsignificant results and low statistical power as “negative,” when they are, in fact, inconclusive

Statistical power is the ability to detect a difference of a given size, if such a difference really exists in the population of interest. In studies with low statistical power, results that are not statistically significant are not negative, they are inconclusive: “The absence of proof is not proof of absence.” Unfortunately, many studies reporting non-statistically significant findings are “under-powered” and are therefore of little value because they do not provide conclusive answers (28).

In some situation, non-statistically significant findings are desirable, as when groups in observational studies are compared with hypothesis tests (P values) at baseline to establish that they are similar. Such comparisons often have low power and therefore may not establish that the groups are, in fact, similar.

Error #18: Not distinguishing between “pragmatic” (effectiveness) and “explanatory” (efficacy) studies when designing and interpreting biomedical research

Explanatory or efficacy studies are done to understand a disease or therapeutic process. Such studies are best done under “ideal” or “laboratory” conditions that allow tight control over patient selection, treatment, and follow up. Such studies may provide insight into biological mechanisms, but they may not be generalizable to clinical practice, where the conditions are not so tightly controlled. For example, a double-masked explanatory study of a diagnostic test may be appropriate for evaluating the scientific basis of the test. However, in practice, doctors are not masked to information about their patients, so the study may not be realistic.

Pragmatic or effectiveness studies are performed to guide decision-making. 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 affected by many, uncontrolled, factors, which limits their explanatory power but that may enhance their application in clinical practice. For example, patients in a pragmatic trial are more likely to have a wide range of personal and clinical characteristics than are patients in an explanatory trial, who must usually meet strict entrance criteria.

Many studies try to take both approaches and, as a result, do neither well (29,30). The results of a study should be interpreted in light of the nature of the question it was designed to investigate (Table 4).

Table 4. Differences between explanatory and pragmatic studies in studies of zinc lozenges for treating the common cold. The pragmatic study was designed to determine whether zinc lozenges would reduce the number and duration of cold symptoms in outpatients and was conducted under conditions faced by consumers of the lozenges. The explanatory study was designed to determine whether zinc is an effective antiviral agent and was conducted under much tighter experimental conditions

Variable	Explanatory	Pragmatic
Diagnosis	positive <i>Rhinovirus</i> culture	3 of 10 symptoms
Evidence of efficacy (outcomes)	weight of nasal mucus, tissue counts	reduced number and duration of symptoms
Setting	in-patient	out-patient
Intervention	controlled by researcher	controlled by patient
Design	masked and placebo-controlled	masked and placebo-controlled
Focus	zinc as an antiviral agent	zinc as a treatment for colds

Error #19: Not reporting results in clinically useful units

The reports below (31,32) all use accurate and accepted outcome measures, but each leaves a different impression of the effectiveness of the drug. *Effort-to-yield measures*, especially the number needed to treat, are more clinically relevant and allow different treatments to be compared on similar terms.

– *Results expressed in absolute terms.* In the Helsinki study of hypercholesterolemic men, after 5 years, 84 of 2,030 patients on placebo (4.1%) had heart attacks, whereas only 56 of 2,051 men treated with gemfibrozil (2.7%) had heart attacks ($P < 0.02$), for an absolute risk reduction of 1.4% (4.1-2.7% = 1.4%).

– *Results expressed in relative terms.* In the Helsinki study of hypercholesterolemic men, after 5 years, 4.1% of the men treated with placebo had heart attacks, whereas only 2.7% treated with gemfibrozil had heart attacks. The difference, 1.4%, represents a 34% relative risk reduction in the incidence of heart attack in the gemfibrozil-treated group (1.4%/4.1% = 34%).

– *Results expressed in an effort-to-yield measure, 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.

– *Results expressed as total cohort mortality rates.* In the Helsinki study, total mortality from cardiac events was 6 in the gemfibrozil group and 10 in the control group, for an absolute risk reduction of 0.2%, a relative risk reduction of 40%, and the need to treat 2,460 men for 1 year to prevent 1 death from heart attack.

Error #20: Confusing statistical significance with clinical importance

In statistics, *small differences between large groups* can be statistically significant but clinically meaningless (12,33). In a study of the time-to-failure for two types of pacemaker leads, a mean difference of 0.25 months over 5 years among thousands of leads is not apt to be clinically important, even if such a difference would have occurred by chance less than 1 time 1,000 ($p < 0.001$).

It is also true that *large differences between small groups* can be clinically important but not statistically significant. In a small study of patients with a terminal condition, if even one patient in the treatment group survives, the survival is clinically important, whether or not the survival rate is statistically different from that of the control group.

Conclusion

The real solution to poor statistical reporting will come when authors learn more about research design and statistics; when statisticians improve their ability to communicate statistics to authors, editors, and readers; when researchers begin to involve statisticians at the beginning of research, not at its end; when manuscript editors begin to understand and to apply statistical reporting guidelines (12,18,19,34-40); when more journals are able to screen more carefully more articles containing statistical analyses; and when readers learn more about how to interpret statistics and begin to expect, if not demand, adequate statistical reporting.

References

- 1 Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation*. 1980; 61:1-7.
- 2 Evans M. Presentation of manuscripts for publication in the British Journal of Surgery. *Br J Surg*. 1989;76:1311-4.
- 3 Mainland D. Chance and the blood count. 1934. *CMAJ*. 1993;148:225-7.
- 4 Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA*. 1966;195:1123-8.
- 5 White SJ. Statistical errors in papers in the British Journal of Psychiatry. *Brit J Psychiatry*. 1979;135:336-42.
- 6 Hemminki E. Quality of reports of clinical trials submitted by the drug industry to the Finnish and Swedish control authorities. *Eur J Clin Pharmacol*. 1981;19:157-65.
- 7 Gore SM, Jones G, Thompson SG. The Lancet's statistical review process: areas for improvement by authors. *Lancet*. 1992;340:100-2.
- 8 George SL. Statistics in medical journals: a survey of current policies and proposals for editors. *Med Pediatric Oncol*. 1985;13:109-12.
- 9 Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne 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: 1191-4; *JAMA*. 2001;285:1987-91.
- 10 Stroup D, Berlin J, Morton S, Olkin I, Williamson GD, Rennie D, et al. Meta-analysis of observational studies in epidemiology. A proposal for reporting. *JAMA*. 2000;283:2008-12.
- 11 Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF, for the Quorum group. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUORUM statement. *Lancet*. 1999;354: 1896-900.
- 12 Lang T, Secic M. How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers. Philadelphia (PA): American College of Physicians; 1997.
- 13 Ehrenberg AS. The problem of numeracy. *Am Statistician*. 1981;286:67-71.
- 14 Murray GD. The task of a statistical referee. *Br J Surg*. 1988;75:664-7.
- 15 Feinstein AR. X and iprP: an improved summary for scientific communication. *J Chronic Dis*. 1987;40:283-8.
- 16 Feinstein AR. Clinical biostatistics XXXVII. Demeaned errors, confidence games, nonplussed minuses, inefficient coefficients, and other statistical disruptions of scientific communication. *Clin Pharm Therapeutics*. 1976;20:617-31.
- 17 Gardner MJ, Altman D. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ*. 1986;292:746-50.
- 18 Bailar JC, Mosteller F. Guidelines for statistical reporting in articles for medical journals. *Ann Intern Med*. 1988;108:266-73.
- 19 DerSimonian R, Charette LJ, McPeck B, Mosteller F. Reporting on methods in clinical trials. *N Engl J Med*. 1982;306:1332-7.
- 20 Cooper GS, Zangwill L. An analysis of the quality of research reports in the Journal of General Internal Medicine. *J Gen Intern Med*. 1989;4:232-6.
- 21 Hampton JR. Presentation and analysis of the results of clinical trials in cardiovascular disease. *BMJ*. 1981;282: 1371-3.
- 22 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;317:426-32.
- 23 Altman DG, Dore CJ. Randomisation and baseline comparisons in clinical trials. *Lancet*. 1990;335:149-53.
- 24 How to read clinical journals: II. To learn about a diagnostic test. *Can Med Assoc J*. 1981;124:703-10.
- 25 Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. *Med Decis Making*. 1987;7:107-14.
- 26 Harris RL. Information graphics: a comprehensive illustrated reference. Oxford: Oxford University Press; 1999.
- 27 Lang T, Talerico C. Improving comprehension: theories and research findings. In: American Medical Writers Association. Selected workshops in biomedical communication, Vol. 2. Bethesda (MD): American Medical Writers Association; 1997.
- 28 Gotzsche PC. Methodology and overt and hidden bias in reports of 196 double-blind trials of nonsteroidal antiinflammatory drugs in rheumatoid arthritis. *Cont Clin Trials*. 1989;10:31-56.
- 29 Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. *J Chron Dis*. 1967;20: 637-48.
- 30 Simon G, Wagner E, Vonkroff M. Cost-effectiveness comparisons using "real world" randomized trials: the case of new antidepressant drugs. *J Clin Epidemiol*. 1995;48:363-73.
- 31 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.
- 32 Brett AS. Treating hypercholesterolemia: how should practicing physicians interpret the published data for patients? *N Engl J Med*. 1989;321:676-80.
- 33 Ellenbaas RM, Ellenbaas JK, Cuddy PG. Evaluating the medical literature, part II: statistical analysis. *Ann Emerg Med*. 1983;12:610-20.
- 34 Altman DG, Gore SM, Gardner MJ, Pocock SJ. Statistical guidelines for contributors to medical journals. *BMJ*. 1983;286:1489-93.
- 35 Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. *Cont Clin Trials*. 1981;2:31-49.

- 36 Gardner MJ, Machin D, Campbell MJ. Use of checklists in assessing the statistical content of medical studies. *BMJ*. 1986;292:810-2.
- 37 Mosteller F, Gilbert JP, McPeck B. Reporting Standards and Research Strategies for Controlled Trials. *Control Clin Trials*. 1980;1:37-58.
- 38 Murray GD. Statistical guidelines for the British Journal of Surgery. *Br J Surg*. 1991;78:782-4.
- 39 Simon R, Wittes RE. Methodologic guidelines for reports of clinical trials. *Cancer Treat Rep*. 1985;69:1-3.
- 40 Zelen M. Guidelines for publishing papers on cancer clinical trials: responsibilities of editors and authors. *J Clin Oncol*. 1983;1:164-9.

Correspondence to:

Tom Lang
Tom Lang Communications
PO Box 1257
Murphys, CA 95247, USA
tomlangcom@aol.com