J Exp Stroke Transl Med (2010) 3(1): 1-7
Society for Experimental Stroke (www.s4es.org)

Common statistical concerns in clinical trials

Scott R. Evans, Ph.D.

Department of Statistics, Harvard University, Boston, MA

Abstract

Statistics are an integral part of clinical trials. Elements of statistics span clinical trial design, data monitoring, analyses, and reporting. A solid understanding of statistical concepts by clinicians improves the comprehension and the resulting quality of clinical trials. This manuscript outlines common statistical concerns in clinical trials that are important for clinicians to understand.

Keywords: p-value; confidence intervals; intent-to-treat; missing data; multiplicity; subgroup analyses; causation

1. Introduction

Statistics are an important aspect of clinical trials. The breadth of statistical issues span the full spectrum of a trial from design and protocol development, data monitoring and conduct during an ongoing trial, data management, data analyses, and reporting of trial results. Statistical concepts can be difficult for non-statisticians to understand. However it is important that clinicians involved with clinical trials understand fundamental statistical issues in order to uphold the integrity of a trial. Knowledge of basic biostatistics and of study design is also important for reading medical journals(Weiss and Samet 1980).

Studies suggest a marked increase in the complexity of statistical methods in the medical literature. (Horton and Switzer 2005) Studies also demonstrate that the current level of medical training in biostatistics is inadequate. (West and Ficalora 2007; Windish *et al* 2007) In a study by West and Ficalora, more than two-thirds of clinician respondents disagreed or strongly disagreed with the statement that "the current level of medical training in biostatistics in medicine is adequate". Clinicians often seem willing to draw conclusions unsupported by the data.(Berwick *et al* 1981) Wulff et.al (Wulff *et al* 1986) reported that statistical knowledge of most doctors is so limited that they cannot be expected to draw the right conclusions from the analyses presented in medical journals. There is evidence that researchers often inappropriately apply statistical methods due to poor understanding of statistical concepts(Altman and Bland 1998) and Glantz(Glantz 1980) suggested that approximately half of the published articles in medical journals that use statistical methods, use them incorrectly.

In this article, I discuss some basic statistical issues that are common concerns in clinical trials.

2. Common statistical concerns in clinical trials

There are several common statistical concerns in clinical trials including poor p-value interpretation, the need for presenting confidence intervals, adherence to the intent-to-treat principle, missing data, multiplicity, subgroup analyses, association vs. causation, appropriate reporting of trial results, probability and Bayesian statistics, and the clinician-statistician interaction and the importance of effective communication.

2.1 Poor p-value Interpretation

The use of statistics in medical journals has increased dramatically over the past few decades. One unfortunate consequence has been a shift of emphasis away from the basic results towards a concentration on hypothesis testing. (Gardner and Altman 1986)

One of the biggest flaws in medical research is the over-reliance on and misinterpretation of the p-value. The p-value is interpreted within the context of a hypothesis test where complimentary hypotheses, a null hypothesis (assumed to be true) and an alternative hypothesis (the claim that researchers wish to prove) are developed.

The p-value is defined as the probability of observing data as or more extreme than the observed data if the null hypothesis was true (note that the p-value is not the probability of a hypothesis being true given the data). If this probability is low (e.g., <0.05) then either: (1) the observation of these data is a rare event, or (2) the null hypothesis is not true. The standard practice is to reject the null hypothesis (in favor of the alternative hypothesis) when the p-value is acceptable low. If the p-value is not acceptably low then there is a failure to reject the null hypothesis.

---

* **Correspondence should be sent to:**

Scott R. Evans, Ph.D., Department of Statistics, Harvard University, 651 Huntington Avenue, FBX 513, Boston, MA, 02115. Phone: 614.432.2998; Fax: 617.432.3163; Email: evans@sdac.harvard.edu

Since the p-value is defined as the probability of observing data as or more extreme than the observed data if the null hypothesis was true, in order to appropriate interpret the p-value, a clear understanding of the null hypothesis is needed. For example, the null hypothesis for a two-sample t-test is the means of two groups are equal. Thus if the null hypothesis is rejected, then one concludes that the means are unqual (the alternative hypothesis). The validity of the t-test is based on an assumption of normality, however this assumption does not always hold. In such a case, statisticians often opt for a Mann-Whitney U test whose validity does not depend on the normality assumption. However the null hypothesis for the Mann-Whitney U test is not the equivalence of means of the two groups but that the ranks of the two groups do not differ (i.e., if one were to rank the outcomes in the two groups combined, that the ranks in one group are not higher or lower than the other group). Since the two-sample t-test and the Mann Whitney U test have different hypotheses, the p-values from these two tests should be interpreted differently.

One may interpret the results of a hypothesis test similarly to the result of a court trial where the null hypothesis is the assumption of innocence and the alternative hypothesis is that the person is guilty. If there is enough evidence to reject the null hypothesis of innocence (i.e., verdict of "guilty"), then one may conclude that there was evidence found to conclude guilt. However if the null hypothesis of innocence is not rejected (i.e., verdict of "not guilty"), then it cannot be said that innocence was proven, only that there was a lack of evidence to conclude guilt. Thus one does not prove the null hypothesis, you only fail to reject it. "Absence of evidence is not evidence of absence." (Altman and Bland 1995)

For example, consider a trial comparing a new therapy vs. placebo. Researchers would like to show that the new therapy is different (better) than placebo (and thus this becomes the alternative hypothesis while its compliment, that the new therapy is not different from placebo is the null hypothesis). Suppose that researchers decide that a 5% false positive error rate is acceptable. Thus if the p-value is <0.05 then the null hypothesis is rejected in favor of the alternative and the conclusion is that the new therapy is different from placebo. If the p-value is not <0.05 then the null hypothesis is not rejected and researchers cannot conclude that the new therapy is different from placebo. Note that "no difference" has not been proven; you are only unable to reject the possibility of "no difference".

The traditional cut-point of 0.05 is arbitrary and p-values are not binary statistics. There is very little difference in the evidence of effect when a p-value is 0.052 vs. 0.048. The 0.05 cut-off is used to control the "false positive" error rate (i.e., to ensure that it is not greater than 5% when the null hypothesis is true). However, researchers can decide for themselves if a 5% false positive error rate is appropriate. If a false positive error is very costly (e.g., would result in a very expensive and invasive therapy being used when effective and safer alternatives are available) then researchers may opt for a 1% false positive rate (i.e., use 0.01 as a cutoff). If a false positive error is not costly, then a 10% (i.e., 0.10 cutoff) could be used. The standard 5% false positive error rate is often used since this is the regulatory hurdle for approval of a new intervention.

The p-value is a function of effect size, sample size, and variability. Larger effect sizes, larger sample sizes, and smaller variation all contribute to smaller p-values. Researchers often incorrectly interpret the p-value as providing direct information about the effect size. P-values do not provide direct information about the magnitude or clinical relevance of the effect. Low p-values (e.g., <0.05) do not imply clinical or practical relevance and high p-values do not imply "no effect". Information about the effect size (or what effect sizes can be ruled out) can only be obtained by constructing confidence intervals.

2.2 Need for confidence intervals

P-values are only one tool for assessing evidence. When reporting the results of a clinical trial, confidence intervals should always be reported to identify effect sizes that can be "ruled out" (i.e., effect sizes that are inconsistent with the data). If a p-value is significant, implying an effect, then the next natural question is "what is the effect?" Confidence intervals directly address this question. If a p-value is not significant, implying that you were not able to rule out the possibility of "no effect", then the next natural question is "what effects could be ruled out?" Confidence intervals again directly address this question. The under-reporting of confidence intervals is a serious flaw in the medical literature.

Confidence intervals are not a *replacement* for p-values but instead should be provided *with* p-values. P-values are still very useful tools particularly when assessing trends and interactions.

Confidence intervals are also frequently misinterpreted. A 95% confidence interval can be thought of as an interval that has 95% probability of covering the parameter of interest (note that this is distinct from a value having 95% probability of falling into an interval). This is to say that if a trial was repeated a very large number of times and each time a 95% confidence interval estimate of the treatment effect was obtained, then 95% of the confidence intervals would cover the true treatment effect. A common misinter-

pretation is that the true treatment effect is more likely to be close to values in the center of the confidence interval. But this is not the case as the center of the interval is no more likely to cover the true treatment effect than values within the interval but away from the center.

2.3 The intent-to-treat (ITT) principle

The ITT principle is a fundamental concept in clinical trials but is frequently misunderstood. The ITT principle essentially states to "analyze as randomized". This means that once a study participant has been randomized then they will be included in the analyses (i.e., the "ITT population") as part of the randomized regimen, regardless of adherence to protocol, study completion, or anything that happens after randomization. Analyzing study participants that did not take their assigned therapy is a difficult concept for researchers to understand. An ITT approach can be considered an evaluation of the treatment "strategy" (i.e., an evaluation of the utility of a treatment in practice) rather than an evaluation of treatment under ideal conditions and perfect adherence.

An alternative to an ITT analyses is a "per protocol" (PP) analyses. The PP analysis is conducted using the PP population, often defined as those study participants that adhere to the randomized therapy and the protocol and have appropriate data for analyses. The PP analysis is appropriate when the goal is to isolate and identify the biological effects of a treatment rather than the utility of the treatment in practice. There are a few key differences between analyses conducted using an ITT population vs. a PP population. The first issue is that randomization is the foundation for statistical inference. Since an ITT analysis is consistent with randomization (e.g., comparisons are indeed randomized comparison), inferences drawn from these analyses are rooted in strong statistical theory. However, PP analyses do not analyze study participants as randomized (e.g., some patients consciously or unconsciously self-select themselves out of the analyses by not adhering to protocol). Thus a PP analysis is not rooted with the same statistical foundation for inference (i.e., treatment comparisons are epidemiological rather than randomized). In a randomized trial, if one observes treatment differences in an ITT analyses, then these differences may be due to differences in treatment or bad luck from randomization (and statistical inference techniques can discriminate between these two). However, if one observes treatment differences in a PP analyses, then these differences may be due to differences in treatment, bad luck from randomization, or from a factor that is causing people not to adhere to protocol (and statistical techniques cannot necessarily isolate the treatment effect). A second key difference is that ITT analyses apply to patients sitting in a clinicians' office waiting to be treated but a PP analyses may not. The PP analysis only compares adherent patients but the future adherence of a patient sitting in a clinicians' office is unknown. However the analysis of treatment strategy (ITT analyses) will apply to the patient. This is attractive since the possibility that the patient will not adhere should be considered when making treatment decisions.

Thus clinical trials should be conducted with ITT principles in mind. Study participants should be followed regardless of adherence or treatment status. Occasionally, treatment may need to be withdrawn when there are concerns for patient safety. However the patient should still be followed on study and planned data should be gathered. It is important to realize the distinction between "off-treatment" and "off-study". Researchers should try to keep study participants on-study regardless of treatment status in accordance with ITT principles. Of course participants can go off study when they no longer consent, for ethical reasons, or if there are safety issues that cannot be adequately addressed by taking the patient off-treatment.

2.4 Missing data

Missing data is one of the biggest threats to the integrity of a clinical trial. Nearly all trials will have missing data but when a trial has substantial missing data, the interpretation of trial results is very challenging and must be viewed with a level of uncertainty. Missing data is usually caused by loss-to-follow-up or patient refusal to participate and provide data. Missing data can be the cause of biased treatment comparisons or estimates of treatment effects. This is because "missingness" is usually not random (e.g., could be related to treatment, or outcome, or both). For example, a treatment-related serious adverse event may cause a patient to discontinue the study and thus outcome data may not be available for this patient at the end of the trial.

Prevention is the first and most effective approach for addressing the missing data problem. Researchers can prevent missing data by designing simple clinical trials (e.g., designing protocols that are easy to comply to; having easy instructions; having patient visits that are not too burdensome, having short, clear case report forms that are easy to complete, etc.) and adhering to the ITT principle (i.e., following patients after randomization regardless of adherence).

Missing data create an obstacle for applying an ITT analysis. The ITT principle states that all randomized study participants should be analyzed, but the data for some participants are missing. The alternative per protocol analyses might only analyze data that are observed. But if patients that will perform poorly, drop out of the trial and only remaining patients are

analyzed, then a distorted view of the therapy will be obtained. In order to comply with the ITT principle, the general approach is to impute data that are missing (i.e., use a strategic "guess" at what the data might have been if observed). For example, if the primary endpoint was binary (i.e., treatment success vs. treatment failure) then an assumption of treatment failure may be imputed and then the data would be analyzed. Sensitivity analyses are then conducted to see how robust the results are to varying assumptions about the missing data. For example, other analyses assuming treatment success may be conducted and the results would be compared. If the results are qualitatively similar then some comfort in interpretation may be gained. But if the results are qualitatively different then interpretation is challenging. If there is a lot of missing data then qualitative differences are likely.

2.5 Multiplicity

Researchers are often interested in testing several hypotheses. Consider a clinical trial designed to compare a new therapy vs. placebo. Researchers may wish to test the effect of the intervention vs. placebo on several outcomes (e.g., a primary outcome and several secondary outcomes). Similarly researchers may wish to test these endpoints in several subgroups of patients (e.g., defined by gender, race, age, or baseline disease status) or at several timepoints during therapy.

Each time a hypothesis test is conducted (i.e., each time a p-value is calculated), there is a chance to make an error (e.g., a false-positive error). For any single test, researchers can control the false positive error rate by deciding the "significance level". For example, it is common to claim that p-values below 0.05 are "significant". This decision rule sets the false positive error rate at 5% for a single test. However if a researcher conducts several tests, then the probability of making at least one false positive error is greater than 5%. The probability of at least one false positive finding increases as the number of tests that are conducted increases. If a researcher conducts 14 hypothesis tests when null hypothesis is true, then the probability of at least one false-positive finding is $[1- (1- 0.05)^{14}]= 0.512$ or 51.2%

Thus it is important for researchers to consider testing only important hypotheses to reduce the possibility of false conclusions. Significant results that are obtained when many tests were conducted without control of the trial-wise false positive error, may need to be validated with independent data.

It is important to report the results of non-significant tests so that significant results can be interpreted within the context of the number of hypothesis tests that were conducted. Researchers can either: (1) clearly report the total number of hypothesis tests, the significance level of each test, and the number of expected false-positive tests by chance or (2) control the false positive error rate with an adjustment to the significance level of each individual test so that the probability of making a trial-wise false positive error is controlled. For example if a researcher plans to conduct two hypothesis tests, then the trial wise error rate could be controlled at 5% by conducting each individual test at 0.025 (rather than 0.05).

2.6 Subgroup analyses

A subgroup analysis is an evaluation of a treatment effect for a specific subset of patients defined by baseline characteristics. Analyses of subgroups that defined by post-baseline characteristics are not advised in clinical trials as such analyses are subject to several types of biases.

Subgroup analyses are subject to the multiplicity problem. Thus subgroup analyses should be conducted selectively. An evaluation of whether a treatment effect varies across subgroups (i.e., treatment effect heterogeneity) should be conducted prior to conducting subgroup analyses. This evaluation is typically conducted via statistical tests for interaction. Only if the treatment effect varies across subgroups should specific subgroup analyses be undertaken. For example, there may be interest in evaluating whether a treatment effect is similar for men vs. women. If the treatment effect varies by gender then subgroup analyses may be undertaken. However, if the treatment effect is not dissimilar then there is no reason to conduct subgroup analyses within each gender.

When evaluating whether the treatment effect varies across subgroups, it is important to clarify the metric. Consider the data in Table 1* displaying the response rate for a new therapy and a control for three age subgroups.

For each treatment group, the response rate increases with age. However there is no interaction (heterogeneity of treatment effects) on the relative risk scale but there is on the absolute scale.

The reporting of subgroup analyses in the literature has generally been poor. Furthermore the results of subgroup analyses are often over-interpreted. First there is often low power to detect effects within subgroups. This is because clinical trials are generally powered to detect overall treatment effects and not necessarily for effects within particular subgroups (where sample sizes are obviously smaller). Furthermore consider a trial that compares a new therapy vs. a control where the primary outcome is a clinical response (vs. no response). Suppose the results for men were 32 of 40 in the new therapy arm responded

vs. 16 of 40 in the control arm. This yields a p-value of <0.01. Suppose the results for women were 4 of 10 in the new therapy arm responded vs. 2 of 10 in the control arm. This yields a p-value of 0.49. Does this imply that the treatment is effective in males but not females? No. Note that the relative risk in each gender is 2. It is only the smaller sample size that leads to the nonsignificant result in females. Note that conducting these subgroup analyses do not address the question of whether the treatment effect varies by gender.

When subgroup analyses are conducted then they should be reported regardless of significance. A forest plot is an effective method for reporting the results of subgroup analyses. The number of subgroup analyses conducted should be transparent so that results can be interpreted within the appropriate context. Subgroup analyses should generally be considered exploratory analyses rather than confirmatory. It is advisable to pre-specify subgroup analyses to avoid "data dredging".

Table 1. Response rate for a new therapy with control for three age subgroups

Age subgroup	New therapy rate	Control therapy rate	Relative risk	Risk difference	Odds Ratio
Young	0.1	0.05	2	0.05	2.11
Medium	0.4	0.2	2	0.2	2.67
Old	0.7	0.35	2	0.35	4.33

2.7 Association vs. causation

A common mistake of clinical researchers is to interpret significant statistical tests of association as causation. Causation is a much stronger concept than association. There are no formal statistical tests for causation (only for association). Although criteria for determining causation are not universal, a conclusion of causation often requires ruling out other possible causes, temporality (demonstrating that the cause precedes the effect), strong association, consistency (repeatability), specificity (causes result in a single effect), biological gradient (monotone dose response), biological plausibility, coherence (consistency with other knowledge), and experimental evidence. Clinical trials try to address the causation issue through the use of randomization and the ITT principle. However even in randomized clinical trials, replication of trial results via other randomized trials is usually needed. This is particularly true for evaluating causes other than randomized treatment. A more common concern is to conclude causation between a non-randomized factor and a trial outcome. Researchers should be very careful about concluding causation without randomization.

2.8 Reporting

Appropriate reporting of clinical trial results is crucial for scientific advancement. Selective reporting is very common and can result in sub-optimal patient care. A common problem in medical research is the under-reporting of negative evidence. If trial results are negative, researchers often elect not to publish these results, perhaps in part because medical journals do not consider the results exciting enough to publish. However, if several trials are conducted to evaluate the effectiveness of a new intervention, and only one trial is positive and furthermore is the only trial that is published, then the medical community is left with a distorted view of the evidence of effectiveness of the new intervention. For these reasons, negative evidence should be reported with equal vigor.

When reporting the results of clinical trials, it is important to report measures of variation along with point estimates of the treatment effect, and confidence intervals. Reporting both relative risk and absolute risk measures, of adverse events for example, are helpful for interpreting the impact of the events. Creative and interpretable data presentation helps to convey the overall message from the trial data. Reporting both benefits and risks (categorized by severity) provides a more complete picture of the effect of a therapy. Providing reference rates (e.g., of no therapy or an alternative therapy) can further help put the results into perspective and aid other clinicians in making treatment decisions.

Researchers can consult the Consolidated Standards of Reporting Trials (CONSORT) Statement, which encompasses various initiatives to alleviate the problems arising from inadequate reporting of randomized controlled trials. The CONSORT Statement is an evidence-based, minimum set of recommendations for reporting randomized clinical trials. It offers a standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting, and aiding their critical appraisal and interpretation. It comprises a 22-item checklist and a flow diagram and is considered an evolving document.

The checklist items focus on reporting how the trial was designed, analyzed, and interpreted; the flow diagram displays the progress of all participants through the trial.

2.9 Probability and Bayesian statistics

P-values are a product of a traditional "frequentist" approach to statistics. A p-value is the probability of observing data as or more extreme than that observed if the null hypothesis is true. In other words it is the probability of the data given a hypothesis being true. Researchers are often more interested in the question "what is the probability that a hypothesis is true given the data?". Traditional frequentist statisticians view this as asking the probability of a fact (i.e., either the hypothesis is true or it is not, and thus the probability is either 0 or 1).

However an alternative statistical approach, Bayesian statistics, allows calculation of the probability of a hypothesis being true given the data. This approach can be more intuitive or appealing to researchers as they wish to know if a particular hypothesis is true. The disadvantage of this approach is that it requires additional assumptions and researchers generally try to move towards fewer assumptions so that results are robust. Bayesian approaches are based on the idea that unknown quantities (e.g., a treatment difference) have probability distributions. The assumptions (called prior distributions in Bayesian terms) often incorporate prior beliefs about the hypothesis. Historical data can be used to help construct the prior distribution. This might be an attractive approach when sound prior knowledge based on reliable data is available. The prior distribution is then updated to a "posterior distribution" based on data collected in the trial. Use of Bayesian statistics has become more common in the design of clinical trials for devices.

A simple example that illustrates the differences between frequentist and Bayesian approaches is in the evaluation of diagnostic tests. Sensitivity (the probability of a diagnostic test being positive when a person is truly diseased) and specificity (the probability of a diagnostic test being negative when a person is truly non-diseased) are examples of frequentist probabilities. However one may wish to know the positive predictive value (the probability that a person is truly diseased given a positive diagnostic test). This probability can be calculated using a Bayesian approach but requires an assumption about the prevalence of the disease in the population from which the patient belongs.

A common dilemma is whether a frequentist or Bayesian approach is "best" for a particular trial. This has caused some divides in the statistical community. However the approaches should not be viewed as competing and in conflict. Instead Bayesian statistics should be seen as another approach or tool to help understand the data.

2.10 Clinician-statistician interaction and communication

One of the keys to the successful design, monitoring, analyses, and reporting of clinical trials is the quality of the interaction between the protocol team clinicians and statisticians. Solid communication is critical. Statisticians need to understand the clinical questions of interest at the deepest level and then develop strategies for answering those questions. They then need to convey the assumptions and limitations of various designs and conduct options to the clinicians. Complicated terminology can often be an obstacle for successful communication. Statisticians do not see patients and do not have a medical background to understand complicated medical terminology. Clinicians can have difficulty understanding complicated statistical terminology. Thus both clinicians and statisticians need to find ways to communicate their ideas in simple terms.

Another potential obstacle to successful communication between statisticians and clinicians is a clear understanding of roles. Statisticians should be viewed as strategists rather than data analysts, programmers, or data managers. Statisticians need to develop a vision of the end of the trial during its design stage. Thus statisticians can be very helpful at preventing problems and should be involved in trial design from the beginning. Statisticians also specialize in study design (e.g., clinical trials vs. epidemiological designs), disease areas, and statistical methods (often determined by the study characteristics typically conducted in disease areas of specialization). Statisticians thus may have limited biological knowledge particularly in a specific disease area that is new to them and need time to develop appropriate knowledge to understand the disease area and related therapies. They may have limited expertise in statistical areas required for particular diseases but often have the flexibility to develop expertise in over time. Thus statisticians may not have answers immediately since they may have to learn new medical areas (e.g., genomics, proteomics, imaging, etc.), the complexities of the data associated with these areas, and the respective statistical methods that are required to analyze such data.

Time constraints can be another obstacle to successful communication between statisticians and clinicians. The demand for statisticians is higher than the supply in clinical trial settings. Statisticians often work on multiple protocols simultaneously. It is also very common for clinicians and other researchers to underestimate the required time of a statistician to help design a trial or analyze data from a trial. Reali-

zation of these issues from clinicians can improve clinician-statistician communication.

3. Summary

There are many places during the life of clinical trials in which errors can take place. (Altman 1998) Fancy statistical methods cannot rescue clinical trial flaws. For this reason, it is optimal to have a statistician that can envision potential obstacles, involved early in the conception of the clinical trial. It is also important that clinicians educate themselves to common statistical issues and concerns when conducting or participating in clinical trials. Perhaps surprisingly, the most important things that clinicians should know about statistics, are not formulas but basic concepts. I have outlined some of the statistical concepts that are commonly misunderstood. Solid understanding of these issues will help to ensure high-quality clinical trials.

Acknowledgement

The author would like to thank Dr. Justin McArthur and Dr. John Griffin for their invitation to participate as part of the ANAs Summer Course for Clinical and Translational Research in the Neurosciences. The author thanks the students and faculty in the course for their helpful feedback. This work was supported in part by Neurologic AIDS Research Consortium (NS32228) and the Statistical and Data Management Center for the AIDS Clinical Trials Group (U01 068634).

References

Altman DG, Bland JM (1995) Absence of Evidence is Not Evidence of Absence. *BMJ* 311:485.

Altman DG (1998) Statistical Reviewing for Medical Journals. *Statist Med* 17:2661-74.

Altman DG, Bland JM (1998) Improving Doctors' Understanding of Statistics. *J R Statist Soc A* 154:223-67.

Berwick D, Fineber HV, Weinstein MC (1981) When Doctors Meet Numbers. *The American Journal of Medicine* 71:991-8.

Gardner MJ, Altman DG (1986) Confidence Intervals Rather than P-Values: Estimation rather than hypothesis testing. *BMJ* 292:746-50.

Glantz SA (1980) How to Detect, Correct and Prevent Errors in the Medical Literature. *Biostatistics* 61:1-7.

Horton NJ, Switzer SS (2005) Statistical Methods in the Journal (letter). *N Engl J Med* 353:1977-9.

Weiss ST, Samet JM (1980) An Assessment of Physician Knowledge of Epidemiology and Biostatistics. *Journal of Medical Education* 55:692-7.

West CP, Ficalora RD (2007) Clinician Attitudes Towards Biostatistics. *Mayo Clin Proc* 82:939-43.

Windish DM, Huot SJ, Free ML (2007) Medicine Residents' Understanding of the Biostatistics and Results in the Medical Literature. *JAMA* 298:1010-22.

Wulff HR, Andersen B, Brandenhoff P, Guttler F (1986) What Do Doctors Know About Statistics? *Statistics in Medicine* 6:3-10.