Title Page.

Avoiding Seven of the Most Common Statistical Errors in Scientific Literature

Jordan M Potter, PharmD Candidate 2019

University of Kentucky

jmpo228@uky.edu

270-217-4160

Alexander H Flannery, PharmD, BCCCP, BCPS

University of Kentucky

Alex.Flannery@uky.edu

502-609-5754

Key Words: statistical errors, interpretation, analysis, biostatistics

Pages: 15

Tables: 2

Figures: 1

Appendices: 0

Financial Disclosures: None reported

Conflicts of Interest: None reported

ABSTRACT

Conducting statistical analysis is a major component of sharing results. Investigators collect enormous amounts of information when testing a hypothesis and translate that data with statistical tests to present their findings. Scholarship of teaching and learning (SoTL) research should adhere to the same rigorous standards required of those dealing with other areas of scholarship and as such, this paper discusses seven common statistical errors likely to be encountered in the scientific literature. While not an all-inclusive list, the following represents areas with a high potential for statistical error and misinterpretation in clinical research and SoTL. More specifically, this paper demonstrates the consequences of assuming normal distributions, examines each component of sample size calculation and its impact on a study, reviews the historical emphasis on p values, examines the components of subgroup analyses, describes methodology to avoid false interpretations of presented results, and identifies resources for ongoing professional development in biostatistics.

INTRODUCTION

In scientific research, conducting statistical analysis is a major component of sharing results. Investigators collect enormous amounts of information when testing a hypothesis and translate that data with statistical tests to present their findings. The accuracy and validity of the results are reviewed rigorously to ensure correctness before being added to the scientific literature. Healthcare providers and educators alike then utilize this literature to make informed decision based on the statistical inferences presented. Unfortunately, an alarming rate of error in statistical analysis exists in the literature resulting in misinformation and misinterpretation. Concern for the implications of these errors has been raised for more than 30 years.[1,2]

Pocock et al reviewed 45 articles published in the *British Medical Journal*, the *Lancet*, and the *New England Journal of Medicine* in 1987 and demonstrated the prevalence of statistical problems even in the most respected journals.[2] In an *Internal Emergency Medicine* article in 2013, Costantino et al countered, stating that of 125 articles reviewed none of the errors would have changed the results of the work. However, in their analysis Costantino et al reports excluding 'more significant items such as study design, outcome, and bias' and still found that 82% of articles analyzed contained errors.[3]

The Accreditation Council for Pharmacy Education has set forth standards for colleges of pharmacy for educating students about biostatistics and study interpretation.[4] Yet a steady rise in published errors exists with a concurrent decline in practitioner confidence.[5-7] In 2009, Bookstaver et al conducted a survey of postgraduate year

1 (PGY1) pharmacy residents consisting of 10 knowledge-based biostatistics and study design questions, as well as Likert-type scales to assess confidence.[7] Of those who responded to at least one knowledge assessment (n=166), the overall mean biostatistics knowledge score was 47.3% ±18.50%.[7] In nearly all fields of healthcare education, low levels of confidence in regard to biostatistics are reported, leading students, residents, practitioners, and educators to rely on authors' abstract conclusions to interpret studies.[5-7] The overall perception of confidence and training in statistical analysis may impact junior and senior faculty alike, regardless of their level of training.

As a service to its authors, the *American Journal of Pharmaceutical Education* (AJPE), published a special article outlining several areas of improvement to increase likelihood of publication in response to commonly noticed flaws in submitted manuscripts.[8] In the 2016 article, Persky and Romanelli provided suggestions to potential authors and reviewers, including recommendations for statistics and p values, adding to previously published guidelines regarding the scholarship of teaching and learning (SoTL).[8-10] This article is a follow-up to those papers to further discuss common errors in statistical analyses of clinical research and SoTL.

SoTL should adhere to the same rigorous standards required of those dealing with other areas of scholarship and as such, this paper discusses seven common statistical errors likely to be encountered in the scientific literature.[9] While not an all-inclusive list, the following represents areas with a high potential for statistical error and misinterpretation in clinical research and SoTL. More specifically, this paper demonstrates the consequences of assuming a normal distribution of data, differentiates the components of a sample size calculation and how components impact conclusions from a research study, challenges the historical emphasis on p values, examines the components of a well-done subgroup analysis, describes methodology to avoid false interpretations of presented results, and identifies resources for ongoing professional development in biostatistics.

COMMON STATISTICAL ERRORS IN THE LITERATURE

As articulated by April McGrath in her 2016 article in *Teaching and Learning Inquiry*, "Being aware of these factors…will put scholars in a better position to evaluate…their own research and that of others."[11]

Missing the distribution. When reporting continuous variables, it is common for researchers to summarize the data with a mathematical mean. Calculating the mean allows for quick comparisons between groups and gives readers a sense of what the 'average' value is for any given variable. The underlying assumption when calculating a mean is that the continuous data is normally distributed. By definition, about 67% of the values of a normal distribution are within ±1 standard deviation of the mean, and about 95% are within ±2 standard deviations.[12]

Unfortunately, most biological data is not normally distributed. Issues then arise when readers see a mean and standard deviation reported and falsely assume that the values from the sample are normally distributed, when in reality they may be skewed due to outliers. To demonstrate this, imagine a group of 5 students have taken an exam and receive scores of 96, 89, 94, 97, and 13. In assessing the exam for appropriateness, you review the average test score (77.8%) and question the difficulty of the exam. This is a simple example, but it is clear to see that these scores are not evenly distributed and suggested a falsely lower average score when in reality the average was much more likely to be in the mid-90s. In addition to misrepresenting descriptive statistics, falsely assuming a normal distribution of continuous variables may yield inaccurate results from statistical analysis. To appropriately analyze non-normally distributed data, researchers must use non-parametric tests, rather than the parametric tests used for normally distributed data.[13]

There are multiple ways to avoid this common error. What is likely the most simple solution to avoid this error is to create a histogram of the data which is being collected. This graphical representation will allow a visual check of normal distribution. If all data points are available, another simple check for normal distribution is a comparison of the mean, median, and mode. If the data is normally distributed, these three values will be equal. As a final possibility to test the normality of data, goodness-of-fit there statistical tests may be run including the Shapiro-Wilk and Kolmogorov-Smirnov tests.[14] Unfortunately, all of these assessments require authors to present all of the data and readers are often left to assume that sample distributions are normal.

Details of the sample size calculation are key to making accurate conclusions. Calculating an appropriate sample size (n) is an integral step in the research process.[15] Deciding the number of people or other experimental units to involve in a study is a seemingly simple task. However, sample size calculations must balance the financial burden of increasing sample size while maintaining the ability to detect an important effect should one exist. In other words, a sample size too large wastes money, time, and resources and a sample size too small may lack the power to answer the study question.[15] There are four components of a sample size calculation: alpha (type I error), beta (type II error), the minimal relevant difference, and the baseline event occurrence. Optimizing the sample size is challenging and as such, the details of the sample size calculation are key to making accurate conclusions as a reader. It is important to note here that a statistical test can either reject or fail to reject a null hypothesis, but can never prove the hypothesis to be true.[13,16] Any claims by the author of 'proving the alternate hypothesis' should raise concern.

The values of alpha and beta are traditionally chosen to be 0.05 and 0.20, respectively. These values, while arbitrary, are conventional in the scientific literature.[13] Any other stated values should be investigated by the reader, if no compelling explanation is given in the text. The value of alpha establishes the acceptable probability of a type I error for the given hypothesis test. Type I error is falsely rejecting the null hypothesis when it is true, therefore, with a traditional alpha of 0.05, the researchers accept a <5% chance of finding a false-positive conclusion.[15,17] Beta, or type II error, is the probability of failing to reject the null hypothesis when it is false.[17] The conventional beta of 0.20 tells readers that the researchers accept a <20% chance of a false-negative conclusion.[15] Power is the complement of beta (1-beta). It is equivalent for investigators to report a power of 0.80 or 80% versus stating a beta of 0.20 or 20%.

The third component of a calculating a sample size is determining the smallest effect of interest. The concept of estimating an effect size prior to collecting the data seems counterintuitive. However, the estimated effect size is critical to the calculation of the number of patients, students, or data points needed for the study. Since sample size is inversely proportional the square of the expected effect size, even small changes in the expected difference have major implications on the sample size required.[15] To demonstrate this, a hypothetical example is provided. Say you created a training program for pharmacy residents to increase publication rates of pharmacy resident projects. As a proud creator, you believe that this program can increase publications by 25%. To test your hypothesis, you randomize residents to either receive the course or complete standard residency training. Completing a sample calculation (Table 1) yields a required sample of 58 residents per group. While recruiting 216 residents may be feasible, estimating that your training will increase publication rates from 50% to 75% is unrealistic and will likely result in failure to reject the null hypothesis of no statistical difference between residents that completed the training and those who did not. Estimating smaller effect sizes allow for greater sensitivity to detect a difference, but requires exponentially more data points (Table 1). As a researcher and a reader, it is important to recognize this balance of producing accurate results, while maintaining a reasonable sample size.

The final component of the sample size calculation is baseline event and variance for the population. Investigators primarily utilize previous studies or pilot studies and their own knowledge and opinions to estimate the difference of baseline event and variance.[17] This has inherent unreliability, however, it is the only way to preemptively estimate effects in new populations.

Sample size calculation is integral in providing adequate power to detect differences between groups. Yet, only 16% of RCTs in major journals had sufficient statistical power (80%) to detect 25% relative difference and only 36% were powered to detect a 50% relative difference.[18] When reviewing positive claims, remember that the details of the sample size calculation are key to making accurate conclusions.

Don't overly rely on p values: other measures/methods may tell you much more. As stated in Perskey and Romanelli's *Insights, Pearls, and Guidance on Successfully Producing and Publishing Educational Research*, "…impactful published papers almost always have statistical analysis," and that means p values are plentiful.[8] P values are considered to be the evidence in favor of random chance as an explanation for a result.[17] This mathematical probability is then compared to the predetermined alpha, which as stated earlier is an arbitrary cutoff. Then, an all-or-none decision is made as to whether the difference between groups was statistically significant or not. P values are especially prone to misinterpretation and as academics it is imperative to recognize that even statistically significant results tell only part of the story.

First, it is necessary to recognize that statistical significance does not measure effect size or relevance and importance of a result. Statistical significance is highly dependent on sample size and even very small effects can appear statistically significant when the sample size is large enough.[17] In the ALBIOS trial, a multicenter randomized controlled trial with 1818 ICU patients with septic shock or severe sepsis, the albumin group had a significantly lower heart rate than those in the crystalloid group (p=.002).[19] Shown only in the appendix, however, are the heart rates of the two groups; 89±20 versus 92±20 beats per minute.[19] As a reader, it is imperative to differentiate statistical significance from a relevant result in the field of study. Similarly, a non-statistically significant result cannot be equated to a statement of no effect. A p value of p=.049 (assuming an alpha of .05) would be considered significant, whereas p=.051would not be statistically significant. In the same manner that a decrease of three beats per minute in heart rate was statistically significant, but not clinically relevant in the ALBIOS trial, a meaningful difference could yield a non-significant result in a different study. A careful assessment of effect size is always necessary and is most easily conducted with review of the confidence intervals. Another erroneous practice is equating the relative size of a p value to being more or less significant or likely to be true. P values do not measure the probability that the hypothesis is true, or the probability that the result was a random effect.[20] Therefore, p=.000001 is no more significant than p=.01, nor is the result any more likely to be true.

Lastly, p values are only credible if all the assumptions from the statistical test were met. Since p values are a direct output from statistical analysis, the underlying assumptions for the respective test must be met for the p value to be accurate. Examples may include using a parametric test for non-normally distributed data or applying a linear regression without first verifying that the relationship between variables is indeed linear.[20] P values alone don't provide evidence with regard to a hypothesis, yet unfortunately, many articles only report p values.[21] Be sure to examine the whole picture when making conclusions from p values. Look for effect sizes and confidence intervals to aid in your assessment. These convey two critical factors that a p value lacks, the magnitude of an effect and the relative importance of an effect.

Multiple hypothesis testing is not without risks. Testing multiple hypotheses may seem benign. Costs associated with conducting research to answer study questions are high and increase constantly. Logically, researchers want to maximize the return on investment and answer as many questions as possible with the data they collect. Technology continues to advance and the wide availability of user-friendly statistical software have made it possible for investigators to run hundreds of tests with ease. Using software in this manner is inappropriate and introduces significant risk of error into the analysis. This is demonstrated in Figure 1, where type I error is plotted as a function of the number of subgroup analyses run. The risk of error in running 100 subgroup analyses may seem obvious, but it is not uncommon for studies to run 20 subgroup analyses. Conducting 20 hypothesis tests increases the probability of type I error from .05 to .64 (assuming α=.05). When reviewing work with multiple subgroup analysis, be aware that the likelihood that at least one of the analyses will be 'statistically significant' by chance alone also increases. Redundant or repetitive statistical analysis of research data conducted in an attempt to find significance is often referred to as 'data dredging.'[22]

Outcomes and subgroups should be determined *a priori*, or before any data collection and analysis has occurred. This is not always possible and often times analyses are performed *post-hoc*, or after data has been collected. In this case, the number of analyses, the selection process for these analyses, and the significant and non-significant results of these analyses should all be disclosed by the author.[16] Authors may also use statistical correction factors, such as the Bonferroni or similar procedures, to adjust the levels of significance accordingly when conducting multiple statistical tests or comparisons.[22] Corrections should be stated by the author so that readers and reviewers can be assured that significance levels have been adjusted appropriately to be more conservative.[23]

Be appropriately skeptical of subgroup analyses. Subgroups analysis can be defined as analysis of intervention effects within subgroups of the sample. Subgroups are beneficial because they can expand the number of questions that may be asked from a limited data set, such as, 'Did the sicker patients benefit from the drug?' or 'Do certain educational strategies benefit certain subgroups of students?' However, the analysis of subgroups should be approached cautiously. The Journal of the American Medical Association published a 'users guide' to subgroup analysis and provide a series of questions that should be asked when assessing the validity of subgroup analyses: Can chance explain the apparent subgroup effect?; Is the effect consistent across studies?; Was the subgroup hypothesis one of a small number of hypotheses developed a priori with direction specified?; and Is there strong preexisting support?[24]

In addition to the inherent risk of error associated with testing multiple hypotheses previously described, readers must be appropriately skeptical of an author's claims based on subgroup analysis. One limitation of subgroup analysis is the possibility that the differences or associations found could be spuriously positive.[12] For example, in 1988 the Second International Study of Infarct Survival (ISIS-2) investigators reported an apparent subgroup effect in the randomized trial comparing streptokinase and aspirin in suspected cases of acute myocardial infarction.[25] "Patients … born under the zodiac signs of Gemini or Libra did not experience the same reduction in vascular mortality attributable to aspirin that patients with other zodiac signs had."[25] Obviously, there is no biological explanation for these statistically significant findings and the investigators made this claim to demonstrate the increased risk of type I error when conducting subgroup analysis.[24,25] Another major disadvantage of subgroup analysis is the risk of making a type II error due to smaller sample size in the subgroup. Readers may interpret the lack of a difference between groups to mean that there is no difference when in reality a difference may exist and the group lacks the statistical power to detect it. Subgroups that are prespecified *a priori*, or at baseline, increase the validity of the analysis. In these instances, stratification and regression techniques may be used to adjust the overall comparison for subgroups.[24]

As a reader, a certain level of skepticism is necessary when reviewing the results of any subgroup analysis. Subgroup analyses are highly susceptible to false positives from multiple comparisons, false negatives due to inadequate power, and often lack transparency in the validity of the conduction.[24] Due to these limitations the information garnered from the analysis of subgroups, even under the best circumstances, should be reserved for generating hypothesis for further study.

Details of modeling are just as important as the rest of the methods section. Even the simplest models, examining the relationship between only two variables, are prone to error.[17] One of the first steps in evaluating a model is assessing its plausibility. Correlation does not confer causation. This adage may seem obvious for variables that appear correlated without any reason for connection. However, when looking at multiple factors, correlations can be deceptive. For instance variable X may demonstrate a correlation statistical significance with variable Y, solely because X and Y are both dependent on a third variable Z.[26] When presenting results, the investigator should provide details on the modeling procedure. These explanations should be thorough and include the methods used to create the model, the assumptions of the data that were made, the limitations of applicability of the model, potential sources of bias, and the methods of validation that were used in the creation of the model.[26] In addition to these criteria, the authors should present an assessment of the fit of the model. Use of a linear regression model is predicated on the fact that the relationship between variables is in fact linear. A variable may yield a statistically significant linear regression despite the underlying relationship not being linear.[17] This should be assessed with an analysis of 'residuals'. Further testing of model fit should be described by goodness-of-fit tests, such as the Hosmer-Lemeshow test.[27] This test measures the statistical significance of any differences between the observed and predicted outcomes over the risk groups. A model that is properly fitted with yield a non-significant difference and suggest that the model is appropriate.[27] This test can be easily misinterpreted. For instance, a retrospective observational study reviewing two different forms of anticoagulation in patients with a traumatic brain injury claimed "an unexpected finding … was the superiority of LMWH over UH with regards to mortality" despite a resulting Hosmer-Lemeshow goodness-of-fit test of p=.066.[28] Readers should also recognize collinearity, two variables put in the model that measure the same thing and predict each other.

Testing too many covariates to fit a model can lead to 'over-fitting' of the model. As a general rule of thumb, for every covariate to be tested, there should be somewhere between ten and twenty outcomes of interest. Similar to 'data dredging' of multiple hypothesis testing, the number of covariates you include in the model increase risk of spurious findings. This is not uncommon to see in the literature, but a striking example is the retrospective analysis of postoperative urinary tract infections by Sultan et al.[29] This study had a large sample size (n=891) of patients who had a Foley catheter placed by either a surgical resident, operating room nurse, or medical resident.[29] Sultan et al claim that medical students need more monitoring because "patients with Foleys placed by medical students were at over 4-fold increased risk of [catheter-associated urinary tract infection]" (OR 4.09, p=.02)

compared to nurses.[29] The issue in making this claim is that only 2.4% of sample experienced the outcome of interest and the authors put 20 factors into a logistic regression model where there were only 22 events. In this instance, parameters are biased toward the extremes when the number of variables approaches the number of events and it becomes nearly impossible to meaningfully conduct a regression. Correcting for these errors would only widen confidence intervals for the studies claim that medical students need more supervision. Taking all of this into account, it is evident that the results of the author's multivariate analysis (OR 4.09, CI 1.22-13.7; OR 2.16, CI 0.75-6.2) simply does not support the conclusion.[29] Like many of the previously listed, the statistically significant findings generated from modeling should be further researched, rather than a basis for conclusive findings.

It's relatively risky...but it doesn't have to be. An important distinction to make when reviewing and interpreting estimated effect size between two groups is the distinction between the odds ratio, relative risk, and absolute risk for a dichotomous outcome.[30] These summary statistics are not interchangeable and result in different values due to the different calculations required. To start, odds are the probability of occurrence of an event or outcome compared to the probability of the event or outcome not occurring. The odds ratio (OR) is then the ratio of the odds of an event or outcome occurring in one group to the odds of the same event or outcome occurring in the other group. Odds ratios are the estimate of effect size for retrospective data.[31] Alternatively, for prospective data results can be expressed in terms of risk, the probability of occurrence of an event or outcome. Typically risk is presented as risk reduction and can be expressed in relative terms as relative risk reduction (RRR) or absolute terms as absolute risk (ARR) and number needed to treat (NNT).[31] RRR estimates the percentage of baseline risk that is reduced as a result of an intervention.[31] The estimate, however, only applies to the study population and has a tendency to over-estimate the beneficial effects in the intervention group.[30]

Despite recommendation from the CONSORT guidelines for randomized controlled trials, authors do not always report both the relative and absolute effect sizes.[32] Calculating risk in absolute terms provides context for the reduction of risk of an outcome by accounting for likelihood that the event will occur in the population at large. Since authors often do not report ARR, it is necessary to calculate it to accurately assess the results of the study. To demonstrate ARR calculation from RRR, consider this hypothetical example provided by Streiner and Norman in their *Chest* commentary looking at the relative risk of death between two groups.[30] In the hypothetical study, 20 individuals died in the intervention group (n=100) and 40 individuals died in the control group (n=100). The relative risk of death in the intervention group is half that of the control group. Then, instead of a sample of 200 individuals,

the hypothetical study was conducted in 20,000 with the same absolute number of deaths, respectively, yielding RR=0.5 again. This clearly demonstrates how only reporting the relative risk can be very misleading. Calculating the absolute risk reduction incorporates the baseline risk associated with the outcome (see citation for calculation).[30] In the first example, ARR=0.2, or a 20% absolute risk reduction in death in the intervention arm. In the second example, with the same relative risk, the absolute risk reduction is 0.002, or a 0.2% absolute reduction of risk of death. Converting these values into NNT, the number of participants required to receive the intervention before preventing one outcome, further demonstrates this difference (NNT=5 vs. NNT=500).

Speaking to results in relative terms is risky. The example in *Chest* clearly demonstrates how relative risk reduction can be a massive overestimate of effect size.[30] Calculating risk and risk reduction in absolute terms, if not presented in the manuscript, is an easy calculation that should always be conducted as a reader to draw appropriate conclusions about the study.

CONCLUSION

The misinterpretation and abuse of statistical analyses has been established many years, yet their presence is nearly ubiquitous. Only about 20% of published manuscripts are free from statistical error.[33] Despite an increasing number of guidance documents/calls for standardization, the key problem remains that there are no interpretations of these concepts that are simultaneously simple, intuitive, correct, and foolproof.[21] As previously mentioned, this article is not an all-inclusive list of statistical errors in the literature. There are, however, many additional resources available to foster knowledge in this area, some of which are given in Table 2. Quality instruction and teaching are the most important goals for higher education, but research efforts help shape best educational practices.[10]

Statistical analyses are a key part of the evidence for making conclusions that impact patient care, educational methods, and allocation of resources. While cumbersome and often complex, formal training in statistics is not a requirement for due diligence as a scholar. This article serves as a general primer so that commonly encountered mistakes and misinterpretations of academic research will not continue to plague authors and readers alike. Recent increases in the demand for evidence based decision making and quality improvement have placed a greater emphasis on producing and evaluating the scientific literature.[8,10] The goal of this work is to provide authors, instructors, and readers of the scientific literature a resource to avoid common pitfalls of statistical analysis in the ever evolving field of health profession education. As academics, a focus on the interpretation and execution is an integral part of practice and should be part of continued professional development.

ACKNOWLEDGMENTS

The authors have no conflicts of interest and no financial disclosures to report.

REFERENCES

1. Glantz SA. Biostatistics: how to detect, correct, and prevent errors in the medical literature. *Circulation.* 1980;61(1):1-7.

2. 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(7):426-432.

3. Costantino G, Casazza G, Cernuschi G, et al. Errors in medical literature: not a question of impact. *Intern Emerg Med.* 2013;8(2):157-160.

4. Accreditation Council for Pharmacy Education Standards 2016. Chicago, IL2015.

5. Windish DM, Huot SJ, Green ML. Medicine residents' understanding of the biostatistics and results in the medical literature. *Jama.* 2007;298(9):1010-1022.

6. Retsas A. Barriers to using research evidence in nursing practice. *Journal of advanced nursing.* 2000;31(3):599-606.

7. Bookstaver PB, Miller AD, Felder TM, Tice DL, Norris LB, Sutton SS. Assessing pharmacy residents' knowledge of biostatistics and research study design. *Ann Pharmacother.* 2012;46(7-8):991-999.

8. Persky AM, Romanelli F. Insights, Pearls, and Guidance on Successfully Producing and Publishing Educational Research. *American Journal of Pharmaceutical Education.* 2016;80(5).

9. Poirier T, Crouch M, MacKinnon G, Mehvar R, Monk-Tutor M. Updated Guidelines for Manuscripts Describing Instructional Design and Assessment: The IDEAS Format. *American Journal of Pharmaceutical Education.* 2009;73(3).

10. McLaughlin JE, Dean MJ, Mumper RJ, Blouin RA, Roth MT. A Roadmap for Educational Research in Pharmacy. *American Journal of Pharmaceutical Education.* 2013;77(10):218.

11. McGrath A. Searching for Significance in the Scholarship of Teaching and Learning and Finding None: Understanding Non-Significant Results. *Teaching & Learning Inquiry: The ISSOTL Journal.* 2016;4(2).

12. Lang T. Twenty statistical errors even you can find in biomedical research articles. *Croatian medical journal.* 2004;45(4):361-370.

13. Bajwa SJ. Basics, common errors and essentials of statistical tools and techniques in anesthesiology research. *J Anaesthesiol Clin Pharmacol.* 2015;31(4):547-553.

14. Sürücü B, Koç E. Assessing the validity of a statistical distribution: some illustrative examples from dermatological research. *Clinical and Experimental Dermatology.* 2008;33(3):239-242.

15. Noordzij M, Tripepi G, Dekker FW, Zoccali C, Tanck MW, Jager KJ. Sample size calculations: basic principles and common pitfalls. *Nephrol Dial Transplant.* 2010;25(5):1388-1393.

16. Thiese MS, Arnold ZC, Walker SD. The misuse and abuse of statistics in biomedical research. *Biochem Med (Zagreb).* 2015;25(1):5-11.

17. Good PI. *Common Errors in Statistics (and How to Avoid Them).* Hoboken, New Jersey2003.

18. Moher D, Dulberg CS, Wells GA. Statistical power, sample size, and their reporting in randomized controlled trials. *Jama.* 1994;272(2):122-124.

19. Caironi P, Tognoni G, Masson S, et al. Albumin replacement in patients with severe sepsis or septic shock. *N Engl J Med.* 2014;370(15):1412-1421.

20. Wasserstein RL, Lazar NA. The ASA's Statement onp-Values: Context, Process, and Purpose. *The American Statistician.* 2016;70(2):129-133.

21. Abbott EF, Serrano VP, Rethlefsen ML, et al. Trends in P Value, Confidence Interval, and Power Analysis Reporting in Health Professions Education Research Reports: A Systematic Appraisal. *Acad Med.* 2018;93(2):314-323.

22. McGaghie WC, Crandall S. Data Analysis and Statistics. *Academic Medicine.* 2001;76(9):936-938.

23. Regehr G. Reporting of Statistical Analyses. *Academic Medicine.* 2001;76(9):938-939.

24. Sun X, Ioannidis JP, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. *JAMA.* 2014;311(4):405-411.

25. Randomized trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2.ISIS-2 (Second International Study of Infarct Survival) Collaborative Group. *J Am Coll Cardiol.* 1988;12(6 Suppl A):3A-13A.

26. Eddy DM, Hollingworth W, Caro JJ, et al. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--7. *Value Health.* 2012;15(6):843-850.

27. Hosmer DW Jr LS, Sturdivant RX. *Applied Logistic Regression, 3rd Edition.* New York, NY: Wiley; 2013.

28. Benjamin E, Recinos G, Aiolfi A, Inaba K, Demetriades D. Pharmacological Thromboembolic Prophylaxis in Traumatic Brain Injuries: Low Molecular Weight Heparin Is Superior to Unfractionated Heparin. *Ann Surg.* 2017;266(3):463-469.

29. Sultan I, Kilic A, Arnaoutakis G, Kilic A. Impact of Foley Catheter Placement by Medical Students on Rates of Postoperative Urinary Tract Infection. *J Am Coll Surg.* 2018.

30. Streiner DL, Norman GR. Mine is bigger than yours: measures of effect size in research. *Chest.* 2012;141(3):595-598.

31. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: Absolute risk reduction, relative risk reduction, and number needed to treat. *Perspect Clin Res.* 2016;7(1):51-53.

32. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ.* 2010;340:c332.

33. Wu S, Jin Z, Wei X, et al. Misuse of statistical methods in 10 leading Chinese medical journals in 1998 and 2008. *ScientificWorldJournal.* 2011;11:2106-2114.

Tables

Table 1. Pearson Chi-square Test for Proportion Difference†

	Difference of 25.0%	Difference of 12.5%	Difference of 6.25%
Group 1 proportion	0.5	0.25	0.25
Group 2 proportion	0.75	0.375	0.3125
Number of sides	2	2	2
Null proportion difference	0	0	0
Computed N per group	58	215	812
Actual power	0.802	0.801	0.800

† Assuming asymptotic normal distribution, normal approximation method for calculation, α=0.05, and β=0.20

Table 2. Resources for Reviewing Biostatistics and Study Design†

Formal Resources
Statistical software resources
ASHP Foundation
ACCP Programs
Several books
Tools for Developing the Thought Process
Editorials
Letters to the Editor
JAMA (Guide to Statistics and Methods)
BMJ (Endgames)
Peer review
Practice and listen
Ask questions

† *See supplemental content for links*