The misuse and abuse of statistics in biomedical research

Matthew S. Thiese*, Zachary C. Arnold, Skyler D. Walker

Rocky Mountain Center for Occupational and Environmental Health, University of Utah, Salt Lake City, Utah, USA

*Corresponding author: Matt.thiese@hsc.utah.edu

Abstract

Statistics are the primary tools for assessing relationships and evaluating study questions. Unfortunately, these tools are often misused, either inadvertently because of ignorance or lack of planning, or conspicuously to achieve a specified result. Data abuses include the incorrect application of statistical tests, lack of transparency and disclosure about decisions that are made, incomplete or incorrect multivariate model building, or exclusion of outliers. Individually, each of these actions may completely invalidate a study, and often studies are victim to more than one offense. Increasingly there are tools and guidance for researchers to look to, including the development of an analysis plan and a series of study specific checklists, in order to prevent or mitigate these offenses.

Key words: a priori; analytical plan; statistical methods; disclosure; transparency; biostatistics

Received: November 30, 2014 *Accepted: January 03, 2015*

Introduction

Utility of biomedical research is a product of appropriate study design, high quality measures, proper selection and application of statistical methods, and correct interpretations of analytical results. Biostatistics are a set of tools that are used to evaluate relationships between results in biomedical research. They are essential for furthering scientific knowledge and understanding. Unfortunately, statistics can be appropriately used, misused and abused, either through concept or application. These concerns have been voiced in the literature since the 1980s, and are still cogent concerns today (1,2). Concerns include poor planning, communication, or understanding of the conceptual framework that the statistical tool is being used to evaluate. Inappropriate study design is often the first actions in research, and has been discussed in the literature (3-5). Problems with data collection and analyses follow directly from study design and are often poorly described in the literature. Recently, additional guidance has been proposed regarding reporting of study methods and results to improve quality and reduce misconduct within biomedical research (6-8).

Computers and statistical software packages have increased the complexity with which data can be analyzed and, consequently, the use of statistics in medical research has also increased. Unfortunately, though the types of errors may have changed, the frequency of statistical misuse has not (9,10). These errors are primarily due to inadequate knowledge and researchers not seeking support from statisticians (9). There is no consensus of optimal methods for biomedical research among biostatisticians, either currently or in the past. There are many differences of opinion in methodological approaches, as exemplified by Frequentist and Bayesianist statistical methodology, or statistical estimation methods proposed by Fischer, Neyman, and Wald among others (11-15). The differences of opinion are so longstanding entrenched that most statistical packages simultaneously pre-

sent multiple estimation results and allow for selection of different methods depending on the type of data and relative strength of each method.

While most misuses of statistics are inadvertent and arise from a lack of knowledge or planning, others may be deliberate decisions in order to achieve a desired statistical result. A recent systematic review and meta-analysis investigating fabrication and falsification of research found that 33.7% of those surveyed admitted to questionable research practices, including modifying results to improve the outcome, questionable interpretation of data, withholding methodological or analytical details, dropping observations or data points from analyses because of a "gut feeling that they were inaccurate" and deceptive or misleading report of design, data or results (3). While it is difficult to discern the differences between the two, the end result is often the same, erroneous relationships and flawed conclusions that are printed and relied upon by others in the field. This report will discuss common misuses and abuses of biostatistics from an epidemiological perspective and provide some guidance on methods to reduce the likelihood for these wrongdoings.

Errors in statistical design

Each statistical test requires certain assumptions to be met and types of data (categorical, continuous, etc.) in order to produce valid results. If these assumptions are not appropriately considered during selection of statistical tests, meaningful errors and misinterpretation of results are possible. At best errors of this nature may be a slight limitation, and at worst may completely invalidate results and their associated conclusions (16). In worst case scenarios, the research study itself may be entirely compromised. It is possible that errors in the application of biostatistics may occur at any or all stages of a study. Furthermore, a single statistical error can be adequate to invalidate any study results (17). Any research investigation can be appropriately planned and performed, however, if incorrect analytical approach is applied the repercussions may be as grave as if the investigation was fundamentally flawed in either design or execution (17).

Errors in the description and presentation of data

Discussions of statistical assumptions are commonly absent from many research articles (18,19). One study reported that nearly 90% of all the published articles evaluated lacked any discussion of statistical assumptions (19). More concerning is that many articles fail to report which statistical tests were utilized during data analysis (20). Only stating that tests were used "where appropriate" is also grossly inadequate, yet commonly done (18,21). Statistical tests are precisely designed for specific types of data, and with the vast array of tests now available, thorough consideration must be given to the assumptions, which guide their selection.

The prevalence of statistical misuse can be explained by the widespread absence of basic statistical knowledge among the medical community (22,23). In a cross-sectional study of faculty and students from colleges of medicine, Gore reported the 53.87% found statistics to be very difficult, 52.9% could not correctly define the meaning of P value, 36.45% ill-defined standard deviation, and 50.97% failed to correctly calculate sample size.

Appropriate treatment of outliers

Outliers are observations beyond what is expected, which may be identified by some statistical variation (e.g. 3 standard deviations above or below the mean) or simple face validity (e.g. body mass index of 65.0) or consensus based upon on clinical reasons. Traditionally, outliers were excluded from analyses because they were thought to be unduly influencing the statistical model, particularly in studies with small sample size (24). While this may be true in some instances, researchers may consciously or unconsciously exclude valid data that don't fit a pre-defined data pattern or hypothesis, therefore committing an error. This may be a simple error that will have minimal impact on results, or it can be a fatal error, which will completely invalidate results. Arguments for identification and omission of outliers are common, however there is little consensus on the appropri-

ate treatment of an outlier. The most comprehensive approach is to analyze data with the extreme observations included and run a second set of analyses excluding these data. Disclosure and complete presentation of both sets of analyses will allow the readers to arrive at their own conclusions regarding relationships between exposures and outcomes. Unfortunately, due to multiple reasons ranging from malfeasance to word count limitations, these analyses are often not performed or presented. Caution should be taken when excluding any data point, and ideally decisions about what should be excluded should be made prior to data collection during the design of the study.

Data transformation and testing for normality

Data may be skewed in biomedical research, requiring different statistical tests. Assessment for the magnitude of skewness through testing of normality is not uniformly performed, and rarely reported. Normality can be assessed graphically or statistically. If data are not normally distributed, either non-parametric analytical techniques should be employed or data need to be transformed to a normal distribution. Mathematically, data transformation is relatively simple, however interpretation of results can be difficult.

Parametric and non-parametric tests

There are numerous types of statistical misuse. Misapplication of nonparametric and parametric tests, failure to apply corrections, and disregard for statistical independence are just a few (25). Over the years, some have attempted to quantify the amount of statistical errors present in published research articles. Four articles have each reported that approximately 50% of articles in medical and dental research contain one or more statistical errors (26-28). It is likely that these percentages are underestimates because many research publications omit or conceal data, rendering post-examination impossible (26).

Similarly, many statistical tests have various versions and applications. Like the tests themselves, the selection of each version must be in accordance with the required assumptions (18). For example, student's t-test is used to compare the means for two sets of continuous sample data. If the data are paired, meaning each observation in one sample has a corresponding observation in the other, then a paired t-test is used. For independent data, there are different forms of the t-test depending upon the variance of the samples. In cases of equal variance, using a two-sample t-test is appropriate. For unequal variance, a modified two-sample test is required. When more than two samples are compared, ANOVA should be utilized. For both t-tests and ANOVA, multiple comparisons may necessitate adjustment through the use of corrections. There is little agreement on when or how to adjust for multiple comparisons (29).

In an examination of the *American Journal of Physiology*, Williams *et al*. discovered that greater than half of all the articles employed unpaired or paired t-tests (19). Of those articles, approximately 17% failed to correctly utilize the t-test for multiple comparisons by modifying the test with either the Bonferroni or some other correction method (19). In the same study, the authors also reported that articles which used the ANOVA test did not specify whether one-way or two-way designs were selected (19). Likewise, Glantz found, while inspecting two journals that approximately half of the articles that used statistics employed the t-test in situations that required a test for multiple comparisons (2).

In some cases, these errors have led to incorrect conclusions (26). More commonly, the conclusions have not been supported by the statistical results. In one publication, 72% or f articles lacked statistical validation for their conclusions (25).

Transparency - disclosure and *a-priori vs. post-hoc* analytical decisions

Research transparency is an increasingly important topic in biomedical research. The decisions that are made, as well as when those decisions are made can play a strong role in the interpretation of study results. An ideal study has one where all potential outcomes are explored prior to data col-

lection, including common elements such as how data are collected, a detailed statistical analysis plan, the alpha level for statistical significance and what tests of association are going to be performed. There are many advantages to be gained having a thorough approach to the design and analysis of the study and documenting the decisions that were made. However, there are unforeseen situations that arise that often require prompt and proper decisions. These can be innocuous such as failure of a data collection tool, to blatant selection bias to achieve a desired outcome. These decisions can easily, and often are, not mentioned in article manuscripts.

Hawthorne effects are potentially found in research studies that conduct observational or interventional study designs with human participants. It is the theory that study participants act differently when they know they are being watched or are aware of their participation in a research study. This change in behavior can commonly be found in audits for companies, it is not uncommon to see an increase in productivity when employees are made aware of an audit. Operant conditioning can also be blamed on Hawthorne effects, leading to results that stray from true behavior or statistical results in studies. Although it is difficult to eliminate any deceptive results deriving from Hawthorne effects, a pre-planned approached can help maintain true and strong statistically significant results (30).

Some of the misuse is because of the nature of research dissemination. There is a publication bias, where statistically significant results are more likely to be published (31). Publication is important for many reasons, including obtaining grants or other funding and achieving tenure in an academic institution. This external pressure to find statistically significant results from research may bias some scientists to select a statistical method that is more likely to yield statistically significant results. Additionally, the inclusion or exclusion of outliers, or even fabrication of data, may be justified in some scientists mind. There is a proposal for a registry of unpublished social science data that has statistically insignificant results (32).

The decision to analyze an exposure-outcome relationship should ideally be made prior to data collection, i.e. *a priori*. When analytical decisions are made *a priori*, the data collection process is more efficient and researchers are much less likely to find spurious relationships. *A priori* analyses are needed for hypothesis testing, and are generally considered the stronger category of analytical decisions. *Post-hoc* or after the fact analyses can be useful in exploring relationships and generating hypotheses. Often *post-hoc* analyses are not focused and include multiple analyses to investigate potential relationships without full consideration for the suspected causal pathway. These can be "fishing" for results where all potential relationships are analyzed. The hazard arises when researchers perform *post-hoc* analyses and report results without disclosing that they are *post-hoc* findings. Based on the alpha level of 0.05, it is likely that by random chance 1 in 20 relationships will be statistically significant but not clinically meaningful. Proper disclosure of how many analyses were performed *post-hoc*, the decision process for how those analyses were selected for evaluation, and both the statistically significant and insignificant results is warranted.

Epidemiological *vs*. biostatistical model building

Multivariate regression is often used to control for confounding and assess for effect modification (33). Often when assessing the relationship between an exposure and outcome there are many potential confounding variables to control for through statistical adjustment in a multivariate model (34). The selection of variables to include in a multivariate model is often more art than science, with little agreement on the selection process, which is often compounded by the complexity of the adjusting variables and theoretical relationships (34). Purely statistical approaches to model building, including forward and backward stepwise building may result in different "final" main effects models, both in relation to variables included and relationships identified (35,36). Reliance on a pre-determined set of rules regarding

stoppage of the model building process can improve this process, and have been proposed since the 1970s (37).

Directed acyclic graphs (DAGs) have been utilized to both mitigate bias and control for confounding factors (38). DAGs hold strong potential for proper model selection, and may be a viable option for proper covariate selection and model creation (39). Although there is no consensus on which method of model building is most appropriate, certain consistencies remain regardless of the model building method used. Proper planning prior to data collection and well before analyses helps to ensure that variables are appropriately collected and analyzed.

Variables to consider as potential confounders

Clinically meaningful relationships identified from past studies.

- Biologically plausible factors based on the purported causal pathway between the exposure and outcome.
- Other factors that the researcher may suspect would confound the exposure-outcome relationship.
- After identifying a comprehensive list of variables that may be effect modifiers or confounders, additional analytical elements need to be considered and decided upon.

Decisions to make prior to data collection

- P-value criteria for potential inclusion in multivariate model.
- Assessment for colinearity of variables and determination of treatment if colinearity is identified.
- P-value for inclusion in final model.
- P-value for inclusion in effect modification (if assessing for effect modification).

Relatively few peer-reviewed articles contain any description of the number of variables collected, criteria for potential inclusion in a multivariate model, type of multivariate model building method used, how many potential variables were included in the model, and how many different assessments were performed.

Interpretation of results

Many statistical packages allow for a multitude of analyses and results, however proper interpretation is key to translation from research to practice. Understanding the implications of committing either a type I or type II error are key. Type I error is the false rejection of the null when the null is true. Conversely, type II error is the false acceptance of the null hypothesis when the null hypothesis is false. Setting alpha levels prior to analyses are important; however, there are many elements that can influence the P-value, including random error, bias and confounding. A P-value of 0.051 compared to an alpha level of 0.05 does not mean that there is no association, moreover it means that this study was not able to detect a statistically significant result. Many researchers would argue that there may in fact be a relationship but the study was not able to detect it. Additionally, committing a type II error can most often be influenced by bias and lack of sufficient statistical power. Complete understanding the implications of potentially committing either of these errors, as well as methods to minimize the likelihood of committing these errors should be achieved prior to beginning a study.

How to combat misuse and abuse of statistics

There is increasing interest in improvement of statistical methods for epidemiological studies. These improvements include consideration and implementation of more rigorous epidemiological and statistical methods, improved transparency and disclosure regarding statistical methods, appropriate interpretation of statistical results and exclusion of data must be explained.

There are two initiatives aimed at biomedical researchers to improve the design, execution and interpretation of biomedical research. One is termed "*Statistical Analyses and Methods in the Published Literature*", or commonly the "SAMPL Guidelines",

and provides detailed guidelines to reporting of statistical methods and analyses by analysis type (40). While relatively new, the SAMPL Guidelines are a valuable resource when designing a study or writing study results. Another initiative is "*Strengthening Analytical Thinking for Observational Studies*" (STRATOS) which aims to provide guidance in the design, execution and interpretation of observational studies (4). Additional resources, including checklists and guidelines have been presented for specific study design types (STROBE, STARD, CONSORT, etc.).

Textbooks and biostatistical journals, including *Biometrika, Statistical Methods in Medical Research, Statistics in Medicine,* and *Journal of the American Statistical Association,* can provide up to date resources for application of statistical analytical plans, interpretation of results, and improvement of statistical methods. Additionally, there are many statistical societies that hold annual meetings that can provide additional instruction, guidance, and insight.

Furthermore, researchers should strive to stay informed regarding the development and application of statistical tests. Statistical tools including splines, multiple imputation, and ordinal regression analyses are becoming increasingly accepted and applied within biomedical research. As new methods are evaluated and accepted in research, there will be an increasing potential for abuse and misuse of these methods.

Perhaps most importantly, researchers should invest adequate time in developing the theoretical construct, whether that is through a DAG or simple listing of exposure measures, outcome measures, and confounders.

Conclusion

There has been, and will likely continue to be misuse and abuse of statistical tools. Through proper planning, application, and disclosure, combined with guidance and tools, hopefully researchers will continue to design, execute and interpret cutting edge biomedical research to further our knowledge and improve health outcomes.

Potential conflict of interest

None declared.

References

1. *Bland JM, Altman DG. Misleading statistics: errors in textbooks, software and manuals. Int J Epidemiol 1988;17:245-7. http://dx.doi.org/10.1093/ije/17.2.245.*
2. *Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. Circulation 1980;61:1-7. http://dx.doi.org/10.1161/01.CIR.61.1.1.*
3. *Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PloS one 2009;4:e5738. http://dx.doi.org/10.1371/journal.pone.0005738.*
4. *Sauerbrei W, Abrahamowicz M, Altman DG, Cessie S, Carpenter J. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. Statistics in medicine 2014. http://dx.doi.org/10.1002/sim.6265.*
5. *Thiese MS. Observational and interventional study design types; an overview. Biochem Med 2014;24:199-210. http://dx.doi.org/10.11613/BM.2014.022.*
6. *Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. BMC Med Res Methodol 2001;1:2. http://dx.doi.org/10.1186/1471-2288-1-2.*
7. *Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1-12. http://dx.doi.org/10.7326/0003-4819-138-1-200301070-00010.*
8. *Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Prev Med 2007;45:247-51. http://dx.doi.org/10.1016/j.ypmed.2007.08.012.*
9. *Ercan I. Misusage of Statistics In Medical Research. Eur J Gen Med 2007;4:128-34.*
10. *Felson DT, Cupples LA, Meenan RF. Misuse of statistical methods in Arthritis and Rheumatism. 1982 versus 1967-68. Arthritis Rheum 1984;27:1018-22. http://dx.doi.org/10.1002/art.1780270908.*
11. *Bayarri MJ, Berger JO. The interplay of Bayesian and frequentist analysis. Statistical Science 2004:58-80.*
12. *Efron B. Bayesians, frequentists, and scientists. J Am Stat Assoc 2005;100:1-5. http://dx.doi.org/10.1198/016214505000000033.*

13. Fisher R. Statistical methods and scientific induction. J R Stat Soc Series B Stat Methodol 1955:69-78.
14. Wald A. Estimation of a parameter when the number of unknown parameters increases indefinitely with the number of observations. Ann Math Stat 1948:220-7. http://dx.doi.org/10.1214/aoms/1177730246.
15. Lenhard J. Models and statistical inference: The controversy between Fisher and Neyman–Pearson. Brit J Phil Sci 2006;57:69-91. http://dx.doi.org/10.1093/bjps/axi152.
16. Jamart J. Statistical tests in medical research. Acta Oncol 1992;31:723-7. http://dx.doi.org/10.3109/02841869209083860.
17. Altman DG. Statistics and ethics in medical research. Misuse of statistics is unethical. BMJ 1980;281:1182-4. http://dx.doi.org/10.1136/bmj.281.6249.1182.
18. Strasak AM, Zaman Q, Pfeiffer KP, Gobel G, Ulmer H. Statistical errors in medical research--a review of common pitfalls. Swiss Med Wkly 2007;137:44-9.
19. Williams JL, Hathaway CA, Kloster KL, Layne BH. Low power, type II errors, and other statistical problems in recent cardiovascular research. Am J Physiol 1997;273:H487-93.
20. Feinstein AR. Clinical biostatistics. XXV. A survey of the statistical procedures in general medical journals. Clin Pharmacol Ther 1974;15:97-107.
21. Welch GE, 2nd, Gabbe SG. Statistics usage in the American Journal of Obstetrics and Gynecology: has anything changed? Am J Obstet Gynecol 2002;186:584-6. http://dx.doi.org/10.1067/mob.2002.122144.
22. Gore A, Kadam Y, Chavan P, Dhumale G. Application of biostatistics in research by teaching faculty and final-year postgraduate students in colleges of modern medicine: A cross-sectional study. Int J Appl Basic Med Res 2012;2:11-6. http://dx.doi.org/10.4103/2229-516X.96792.
23. Wulff HR, Andersen B, Brandenhoff P, Guttler F. What do doctors know about statistics? Stat Med 1987;6:3-10. http://dx.doi.org/10.1002/sim.4780060103.
24. Hawkins DM. Identification of outliers: Springer; 1980. http://dx.doi.org/10.1007/978-94-015-3994-4.
25. Schor S, Karten I. Statistical evaluation of medical journal manuscripts. JAMA 1966;195:1123-8. http://dx.doi.org/10.1001/jama.1966.03100130097026.
26. Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. BMJ 1977;1:85-7. http://dx.doi.org/10.1136/bmj.1.6053.85.
27. Kim JS, Kim DK, Hong SJ. Assessment of errors and misused statistics in dental research. Int Dent J 2011;61:163-7. http://dx.doi.org/10.1111/j.1875-595X.2011.00037.x.
28. White SJ. Statistical errors in papers in the British Journal of Psychiatry. Br J Psychiatry 1979;135:336-42. http://dx.doi.org/10.1192/bjp.135.4.336.
29. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;316:1236-8. http://dx.doi.org/10.1136/bmj.316.7139.1236.
30. Holden JD. Hawthorne effects and research into professional practice. J Eval Clin Pract 2001;7:65-70. http://dx.doi.org/10.1046/j.1365-2753.2001.00280.x.
31. Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias—an updated review. PloS one 2013;8:e66844. http://dx.doi.org/10.1371/journal.pone.0066844.
32. Franco A, Malhotra N, Simonovits G. Publication Bias in the Social Sciences: Unlocking the File Drawer. Lancet 1991;337:867-72.
33. Hair JF. Multivariate data analysis. 2009.
34. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 2007;26:5512-28. http://dx.doi.org/10.1002/sim.3148.
35. Derksen S, Keselman H. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 1992;45:265-82. http://dx.doi.org/10.1111/j. 2044-8317.1992.tb00992.x.
36. Wiegand RE. Performance of using multiple stepwise algorithms for variable selection. Stat Med 2010;29:1647-59.
37. Bendel RB, Afifi AA. Comparison of stopping rules in forward "stepwise" regression. J Am Stat Assoc 1977;72:46-53.
38. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol 2008;8:70. http://dx.doi.org/10.1186/1471-2288-8-70.
39. Weng H-Y, Hsueh Y-H, Messam LLM, Hertz-Picciotto I. Methods of covariate selection: directed acyclic graphs and the change-in-estimate procedure. Am J Epidemiol 2009:kwp035. http://dx.doi.org/10.1093/aje/kwp035.
40. Lang TA, Altman DG. Basic Statistical Reporting for Articles Published in Biomedical Journals: The "Statistical Analyses and Methods in the Published Literature" or The SAMPL Guidelines". Science Editors' Handbook, European Association of Science Editors. 2013.