African Journal of Biotechnology Vol. 7 (25), pp. 4819-4826, 29 December, 2008
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB08.052
ISSN 1684–5315 © 2008 Academic Journals

Review

Statistical problems in medical research

UM Okeh

Department of Industrial Mathematics and Applied Statistics, Ebonyi State University, Nigeria.
E-mail: umokeh1@yahoo.com. Tel: +2348034304278.

Accepted 27 October, 2008

Many medical specialties have reviewed the drawbacks of statistical methods in medical diagnosis in specialized areas in their journals. To my knowledge this has not been done in general practice. Given the main role of a general practitioner as a biostatistician, it would be of interest to enumerate statistical problems in assessing methods of medical diagnosis in general terms. In conducting and reporting of medical research, there are some common problems in using statistical methodology which may result in invalid inferences being made. This paper is aimed to highlight to inexperienced statisticians, medical practitioners and personnel as well as other non-statistician some of the common statistical problem countered when using statistics to interpret data in medical research. And also comments on good practices to avoid some of these problems.

Key words: Statistical methodology, medical research, diagnostic test.

INTRODUCTION

Statistical procedures in medical diagnosis allow inferences to be extended beyond study subjects to the study population e.g. future patients (Altman and Bland, 1998). Furthermore, statistical methods allow use of information in an objective way and take into account the sampling variability. Thus statistical methodology is an integral part of modern medical research (Young, 1999). However there are many drawbacks that are encountered when using these statistical procedures. Statistics is a huge discipline with different paradigms, schools of thought and alternatives methodologies such that sometimes rationales for choosing one method over the other can be confusing. Inappropriate statistical procedures are sometimes used when using multiples comparisons of several independent groups, use of statistical power, inadequate analysis of reported measurement studies, validation of diagnostic test, statistical utility of multiple diagnostic tests, in designing a study, selecting the types of variables and the distributions of variable.

The basic aim of the analysis and the study design determine the appropriate statistical analysis procedure to be used (Hayran, 2002). The assumptions underlying various statistical and mathematical methods can easily be neglected or violated. In this paper, some of the common statistical drawbacks discussed. This is basically intended for inexperienced statisticians or medical practitioners involved in research.

"Critical reviewers of the biomedical literature have consistently found that about half the articles that used statistical methods did so incorrectly" (Glantz, 1980).

"Good research deserves to be presented well, and good presentation is as much a part of the research as the collection and analysis of the data. We recognize good writing when we see it, let us also recognize that science has the right to be written well" (Evans, 1989).

STATISTICAL UTILITY OF MULTIPLE DIAGNOSTIC TESTS

For a silent disease and a major health problem in our aging society called osteoporosis where diagnostic and risk assessment rely on diagnostic tests, the accuracy and cost of these tests may greatly and diagnostic results based on current WHO criteria are inconsistent. With such a variety of diagnostic tests and measurement sides, clinicians must determine the best diagnostic strategy for specific patient populations, both for screening and for selecting and monitoring treatment. Because of the lack of appropriate statistical tools to assess diagnostic utility of combining multiple tests in their accuracy in predicting osteoporotic fractures and cost-effectiveness, it is difficult to identify the optimum combination of tests. Effort should be geared towards evolving statistical methods for evaluating the utilities of combinations of multiple diagnostic tests performed in se-

quence or in parallel. It is also possible to improve diagnostic consistency based on statistical principles. Furthermore, one can also apply new statistical methods to existing epidemiological databases to identify the optimal combination of diagnostic tests for osteoporosis and most uniform criteria for consistent diagnosis (Lu et al., 2006).

FAST VALIDATION OF DIAGNOSTIC TEST

This may require fast evaluation of a diagnostic test for prognostic prediction. To formally establish a method for prognostic prediction, we need to conduct a prospective study to follow a group of patients. For rare prognostic outcomes, such studies can be costly and take for a long time. The cost of such studies will exclude low cost techniques because it is no possible to recover the costs of a longitudinal prospective study. Also, the change in technology is accelerated. New diagnostic methods can be moving targets. It is very possible that the technique will have been outdated or irrelevant when a long-term validation study is finished. One can improve statistical designs that combine a cross-sectional case-control study with a short-term follow-up study. The new parameters will be measured with those established prognostic predictors for patients with and without prognostic outcomes in the cross-sectional study. Similar parameters and their changes will also be measured in the short-term follow-up study. Under some mild statistical assumptions and improved analysis methods, we hope that we can correctly estimate the prospective prediction power of new parameters within a relatively short-time.

PROBLEMS OF STUDY DESIGN

One very vital thing is to make the correct choose of study design which will enable one to answer the research question in a cost effective way. Often times, study design influences its cost through the sample size number of arms, number of follow-up visits per study participant and the amount of testing to be done, among other factors. Neglecting statistical advice on study design is one of the commonly witnessed “drawbacks” in research.

Apart form choosing the most effective study design and sample size calculations, this stage also involves specifications of main hypotheses, outcomes, potential confounding or risk factors; and for randomized controlled trials, defining randomization and blinding procedures. It is important therefore to highlight some errors encountered involving sampling plan, sampling size calculations, randomizations and sample pooling.

Planning for sampling

Avoidance of bias requires that sampling plan has to be properly done. For instance in assessing prevalence of HIV, a sample of certain group e,g. pregnant women cannot represent the general population as pregnant women are highly sexually active individuals. Also non-response in behavioral health studies can easily be due to self-selection which introduces selection bias.

CALCULATIONS INVOLVING SAMPLE SIZE

Normally, sample size is calculated to obtain estimates of desired precision or to discover any existing effect, for instance, a minimum detectable difference between two treatments. If the sample is smaller than necessary, then enough power for statistical conclusions would not be available. Obviously unnecessarily larger samples would require ore resources than could be justified by the gain in precision or power to detect the difference.

The following vital points should be highly considered when calculating a sample size for a study.

1. The statistical model or test e.g. paired to t-test or two independent sample t-tests to be used for analysis.
2. The level of accuracy of the estimate or detectable difference required.
3. The variations in the population ie how individual data points vary around the expected value
4. The type of sampling technique used e.g. systematic sampling, stratified sampling, etc.

Most often we encounter drop-outs or loss to follow-up in a cohort study. This can commonly be witnessed in transient populations like migrant groups or job seekers moving from place to place. When the number of study participant who are lost to follow-up is large, it may lead to a substantial reduction in the sample size and subsequently loss of power to test the hypothesis or loss of precision in estimating the size of an effect. So in calculating sample size, it is necessary to have an estimate of the dropout rate. This rate should be factored in the calculation of the sample size so that the final sample size is more than the required effective sample size. This will ensure that if the number of participants lost to follow-up during the study is not more than the anticipated drop- out rate, the study will still have the required power or precision.

Sample pooling

This may sometimes be described as laboratory assays. When carrying out a study that requires an expensive assay to detect, the presence of an uncommon characteristic in blood samples it may be advantageous to pool samples in order to reduce the number of tests performed and hence the cost. Such sample pooling is only cost-effective if the probability of a positive test is small. In this case, statistical knowledge is useful to calculate the most

effective number of samples to be pooled, and estimate the expected number of vials required for follow-up on positive tests. Sample pooling has a common mistake of not considering the probability of samples testing positive and calculating the expected number of tests to be done, which may result in resting more samples than necessary. The cost saving in terms of the assay need to be matched by the drawing of a sample of sufficient amount to permit both individual testing when the pooled sample is positive and contribution to a pooled sample.

Randomization

The primary aim of experimental clinical studies is usually to compare effects of treatment regimens. Therefore, if the groups differ in other characteristics apart from the treatment regimen, the comparisons may be biased if prognosis is related to some of these factors. It is therefore, important that groups are as balanced in terms of all other factors (both known and unknown) as possible. Unknown factors cannot be easily adjusted at analysis stage unlike for known potential confounders. Randomization is one of the statistical tools used to ensue that treatment groups are balanced.

If randomization is done correctly, any imbalances between groups are due to chance alone. Randomization using blocks ensures that the numbers of participants are balanced between groups. Blocking is particularly necessary in small studies because simple randomization can lead to imbalance in the number of participants in the trials arms which could reduce the power of a study (Piantadosi, 1997; Mathews, 2000). However, care is needed when deciding on the length of the blocks so that they are short enough to balance the groups but not too long such that investigators are able to predict the assignment of an individual treatment. Other forms of randomization used include stratification and minimization techniques to ensure balancing with respect to known prognostic factors (Piantadosi, 1997; Mathews, 2000).

LOW STATISTICAL POWER

Statistical power is the probability of getting a statistically significant result if there is a biologically real effect in the population being studied. Type 1 error is the probability of rejecting the mill hypothesis falsely. Its counterpart is the type 2 error (termed β), the probability of accepting the null hypothesis falsely, that is, of rejecting the fact that there is a difference between the two groups.

The power of a test is calculated as 1-β, a measure of the ability to detect the real difference if it is there. If the sample size is too small, then it may not be possible to establish the significance of a given difference but that does not mean that the difference is not there. Power analysis allows us to be certain that we have looked hard enough for the difference.

One of the first studies to draw attention to these problems in medicine was that by Freiman et al. (1978) who examined 71 randomized trials that compared the effects of two drugs or treatments and concluded that there was no statistically significant difference between their effects. They showed that, in many of those studies, the actual responses were quite large, but because the sample sizes were too small there was a greater than 10% chance of missing a true 25% therapeutic improvement in 67 of the trials and a true 50% therapeutic improvement in 50 trials. In some instances, this led the investigators to discontinue studying the new treatment and to conclude that it was of no benefit, clearly this is an undesirable outcome; a 25% improvement in the cure rate in any disease would be very welcome. Meanwhile, there are several statistical programs for power analysis that are either free or for purchase that have been evaluated by Thomas and Krebs (1997) and most of these can be found on an excellent web site resources or power analysis (US Geological survey, patuxent Wild life Research centre). The calculations are best done a priori, that is, in planning the study and before starting it, but they can also be done post hoc in determining the power of a study that has been completed.

INCORRECT USE OF MULTIPLE COMPARISONS OF SEVERAL INDEPENDENT GROUPS

Glantz (1980) identified these as among the most frequent errors of statistical analysis. Wallenstein et al. (1980) dealt very effectively with the problem, but it appears that the pendulum has swung too far in the other direction, that is, that correction for multiplicity is sometimes used when it is not needed. People have a great deal of difficulty in deciding when corrections multiplicity is needed, and there are even times when statisticians disagree (Dunnett, 1970). Nevertheless, the general principles are straightforward. In addition, we would like to describe some other analyses that can be used in certain circumstances when the issues of multiple comparisons arise.

Let us illustrate that repeated t-tests shift the probability from a single test. For a simple example, consider that you are trying to decide whether to go home daily or stay and work for another 2 h. To make the decision you will toss a coin. If it lands with heads up, you will go home early, if not, you will stay and work. You toss the coin and it comes up tails. You toss it again, and again it comes up tails. You continue tossing until it comes up heads, so you pack up and go home early. Obviously, at the first toss, there is a 50:50 chance of heads coming up but as you continue tossing, there will eventually be near certainty that a head will appear, the changes of getting 10 trials in a row are 0.000976525.

For a more detailed discussion, this work can do no

better than use an explanation given by Tukey (1977), with a normal distribution. Set the probability of falsely rejecting the null hypothesis (which we know to be true) at 0.05. Therefore, the probability of correctly accepting the null hypothesis is 1-0.05-0.95. Now draw two more groups at random from the same populations, and once again there is a probability of 0.05 of falsely rejecting the null hypothesis and 0.95 of correctly accepting the null hypothesis. Now, what happens if it is stated that the null hypothesis will be rejected if either of the two sets show a significant difference? The probability of correctly accepting the null hypothesis for both sets is the product of the two probabilities: 0.95x0.95=0.9025.

Therefore, the probability of falsely rejecting the null hypothesis is 1-0.9025 = 0.0975. In other words, by giving oneself two chances to reject the null hypothesis, one would have almost doubled the chances of falsely rejecting it. You continue to draw pairs of groups at random from the parent populations, hypothesis increases steadily. Therefore, as the number of t-tests increases, the risk of a type 1 error increases, even though for each individual t-test the risk remains at 0.05. One of the ways of reducing the type1 error is to divide the probability of marking a type 1 error by the number of comparisons (t tests). This ratio remains close to the conventional 0.05 value as may be shown. This is the basis of the Bonferroni correction. For further reading see, Creasy et al. (1972).

MEASUREMENT REPORTING WITH IRRELEVANT PRECISION

Rounding members to two significant digits improves communication (Ehrenberg, 1981). In the sentence below the final population size is about three times the initial population size for both he women and the men, but this fact is only apparent after rounding:

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

The number of women rose from about 30,000 to 94,000 and the number of men rose from about 13,000 to 36,000. Many numbers do not need to be reported with full precision. If a patient weighs 60 kg, reporting the weight as 60.18 kg adds only confusion, even if the measurement was that precise. For the same reason, the smallest P value that needs be reported is P<0.001.

The incorrect application of descriptive statistics: continuous data has means and standard deviation as the most common descriptive statistics. They describe only a “normal” distribution of values. By definition, about 68% of the values of a normal distribution are within plus or minus 1 standard deviation of the means, about 95% are within plus or minus 2 standard deviations, and about 99% are within plus or minus 3 standard deviation. In markedly non-normal distributions, these relationships are no longer true, so that means standard deviation do not communicate the shape of the distribution well. Instead, other measures like median, range, interquartile range are recommended (Murrary, 1988).

Even though mean and standard deviation can be calculated from as few as two data points, these statistics may not describe small samples well. In addition, most biological data are not normally distributed (Feinstein, and Ipr, 1987). For these reasons, the median and range or interquartile range should probably be far more common in the medical literature than the mean and standard deviation.

THE USE OF STANDARD ERROR OF THE MEAN (SEM) AS A DESCRIPTIVE STATISTIC OR AS A MEASURE OF PRECISION FOR AN ESTIMATE

The mean and standard deviation describe the center and variability of normal distribution of a characteristic for a sample. The mean and standard error of the mean (SEM) however, are an estimate (the SEM) for a characteristic of a population. However, the SEM is always smaller than the standard derivation, so it is sometimes reported instead of the standard derivation to make the measurements look move precise (Feinstein, 1976). Although the SEM is a measure of precision for an estimate (1 SEM on either side of the mean is essentially a 68% confidence interval), the preferred measure of precision in medicine is the 95% confidence interval (Gardner and Altman, 1986). The mean and standard deviations are the preferred summary statistics for (normally distributed) data, and the mean and 95% confidence interval are referred for reporting an estimate and its measure of precision.

INTERPRETATION OF P VALUES FOR RESULT

P values are often misinterpreted. Its limitations are not considered even if it is interpreted correctly. For mean results, report the absolute difference between groups (relative or percent differences can be misleading) and 95% confidence interval for the difference instead of or in addition to, p values. The sentences below go from poor to good reporting:

- “The effect of the drug was statistically significant”. This sentence does not indicate the size of the effect, whether the effect is clinically important, or how statistically significant the effect is. Some readers would interpret “statistically signification” in this case to mean that the study supports the use of the drug.
- “The effect of the drug on lowering diastolic blood pressure was statistically significant (P<0.05). Here the size of the drop is not given, so its clinical importance is not known. Also, P could be 0.049; statistically significant (at the 0.05 level) but so close to 0.05 that it should probably be interpreted similarly to a p value of say, 0.51, which is not statistically significant.

The use of an arbitrary cup point, such as 0.05, to distinguish between "significant and "non significant" results is one of the problems of interpreting P values. When a study produces a confidence interval in which all the values are clinically important, the intervention is much more likely to be clinically effective. If none of the values in the interval are clinically important, the intervention is likely to be infective. If only some of the values are clinically important, the study probably did not enroll enough patients.

USING GRAPHICAL TOOLS

Figures and tables should not be used to "store" data i.e just throwing software output in the table graph which does not aid the interpretation. Good statistical graphical and text tools have to be used for reporting summarized data and information in a useful and non-misleading manner and to aid interpretation of the result.

VIOLATION OF THE ASSUMPTIONS OF THE STATISTICAL TESTS

It is frequent and common experience that a researcher will apply a statistical method to a set of data without thoroughly checking that the assumption of the method is valid (Okeh and Ugwu, 2008). This often leads to achieving wrong result. For this reason both the name of the test and a statement that its assumptions were met should be included in reporting every statistical analysis. For example: "the data were approximately normally distributed and thus did not violate the assumptions of the t-test".

The most common problems are:

- Using parametric test when the data are not normally distributed (skewed).
- Using tests of independent samples are paired samples, which require tests for paired data. Again students t-test often used when a paired t-test is required.

MISSING DATA

Missing of data can be common in some variables e.g CD4 count, lead levels in the body or behavioral characteristics: smoking status and drinking habits. Missing data could be due to a whole range of reasons eg limited precision of the recording machine or interviewee's non-response. Missing data can be non-random and ignoring it in the analysis introduces bias. An example of non-random missing data is levels of alcohol consumption where alcoholics are likely to having missing data due to non-response.

Another form of missing data is loss follow-up e.g. in a study of HIV infected individuals where the outcome is morbidity or mortality, patients may be lost to follow-up if they were too sick to come for follow-up visits or died and the researcher was unable to trace them and therefore coded as missing. This will cause bias and needs to be considered when analyzing the data as the degree of missing depends on the outcome.

CHOICE OF MODEL

It is very important to choose an aspect of the study design to model. For instance, ignoring some features like dependence among observations can result in inefficient estimators (Poirier et al., 2003). Dependence occurs when data are collected from an individual over a period of time or from a group of people who are in clusters e.g. children in a classroom and paired data. Ignoring dependence gives invalid inferences due to underestimating of standard errors. For example, use of two sample t-text for paired data is clearly inappropriate.

Model choice also involves choosing the functional form of the relationship between the response and explanatory variables. All assumptions should be evaluated before using a model to ensure that valid inferences are made. Before selecting a model, researchers should evaluate the assumptions implied by the model against the data and prior in formation.

Another aspect of model choice is variable categorization. Categorization of continuous variables is very common in order to simplify the analysis. However, this may result in loss of information. Therefore categorization should be done only when necessary (Royston et al., 2006).

STATISTICAL SOFTWARE PROGRAM

This program with graphical user interface has brought many advantages but also problems. Menu-driven software encourages or permits blind and incorrect use of statistical methods. With robust software, some of the errors can easily go unnoticed or ignored and this has increased the danger of applying inappropriate analysis methods. It is also common to have software output including some irrelevant statistics under specific model assumptions.

DEFINITION OF NORMAL/ABNORMAL IN REPORTING DIAGNOSTIC TEST RESULT

The importance of either a positive or a negative diagnostic test result depends on how "normal" and "abnormal" are defined. In fact, "normal" has at least six definitions in medicine (see, How to read Clinical Journals, 1981).

A diagnostic definition of normal

This is based on the range of measurements over which the disease is absent and beyond which it is likely to be present. Such a definition of normal is desirable because it is clinically useful.

A therapeutic definition of normal

This is based on the range of measurements over which a therapy is not indicated and beyond which it is beneficial. Again, this definition is clinically useful.

Other definition includes risk factor, statistical, percentile, and social definitions of normal which are perhaps less useful for patient care through there common.

NOT EXPLAINING HOW UNCERTAIN EVALUATION OF A SCREENING TEST RESULT WERE TREATED

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

Uncertain test results may be one of three types (Simel et al., 1987):

Intermediate results

These fall between a negative result and a positive result. In a tissue test based on the presence of cells that stain blue, “bluish” cells that are neither unstained nor the required shade of blue might be considered intermediate results.

Indeterminate results

These are results that indicate neither a positive nor a negative finding. For example, responses on a psychological test may not determine whether the respondent is or is not alcohol-dependent.

Uninterpretable results

These are produced when a test is not conducted according to specified performance standards. Glucose levels from patients who did not fast overnight may be uninterpretable, for example.

How such results were counted when calculating sensitivity and specificity should be reported. Test characteristics will vary, depending on whether the results are counted as positive or negative or were not counted all, which is often the case. The standard 2 x 2 table for computing diagnostic sensitivity and specificity does not include rows and columns for uncertain results. Even a highly sensitive or specific test may be of little value if the results are uncertain much of the time.

WRONG PLACEMENT OF UNITS OF OBSERVATION IN REPORTING AND INTERPRETING

What is actually being studied is the unit of observation but if the unit is any other thing expects the patient, problem arises. For instance in a study of 50 eyes, how many patients are involved? What does a 50% success rate mean?

If the unit of observation is the heart attack, a study of 18 heart attacks among 1,000 people has a sample size of 18, not 1,000. The fact that 18 of 1000 people had heart attacks may be important, but there are still only 18 heart attacks to study.

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

INABILITY TO DISTINGUISH BETWEEN “PRAGMATIC” (EFFECTIVENESS) AND “EXPLANATORY” (EFFICACY) STUDIES WHEN DESIGNING AND INTERPRETING MEDICAL RESEARCH

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

Pragmatic or effectiveness studies are performed to guide decision-making. These studies are usually conducted under “normal” conditions that reflect the circumstances under which medical care is usually provided. The results of such studies may be affected by many, uncontrolled, factors, which limit their explanatory power but that may enhance their application in clinical practice.

For instance, patients in a pragmatic trail are more likely to have a wide range of personal and clinical characteristics than are patients in an explanatory trial, who must usually meet strict entrance criteria.

Many studies try to take both approaches and as a results, do neither well (Schwartz and Lellouch, 1967; Simon et al., 1995). The results of a study should be interpreted in the light of the nature of the question it was designed to investigate.

FAILURE TO REPORT MEDICAL RESULTS IN CLINICALLY USEFUL UNITS

The reports below (Guyatt et al., 1994; Brett, 1989) all use accurate and accepted outcome measures, but each leaves a different impression of the effectiveness of the drug. Effort-to-yield measures, especially the number needed to treat, are more clinically relevant and allow different treatments to be compared on similar terms.

Result expressed as total cohort mortality rates

In the Helsinki study, total mortality from cardiac events was 6 in the gemfibrozil group and 10 in the control group, for an absolute risk reduction of 0.2%, a relative risk reduction of 40%, and the need to treat 2,4600 men for 1 year to prevent 1 death from hearth attack.

Results expressed in absolute terms

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

Results expressed in another effort-to-yield measure

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

Results expressed in an effort-to-yield measure, the number needed to treat

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

MISUNDERSTANDING STATISTICAL SIGNIFICANCE FOR CLINICAL IMPORTANCE

In statistics, small differences between large groups can be statistically significant but clinically meaningless (Lang and Secic, 1997). In a study of the time-to-failure for two types of pacemaker leads, a mean difference of 0.25 months over 5 years among thousands of leads is not apt to be clinically importance, even if such a difference would have occurred by chance less than 1 time 1,000 ($p<0.001$).

It is also true that large differences between small groups can be clinically important but not statistically significant.

In a small study of patients with a terminal condition, if even one patient in the treatment groups survives, the survival is clinically important, whether or not the survival rate is statistically different from that of the control groups.

REPORTING PROBLEM

Arguably, errors conducted during analysis or reporting stage usually have relatively low gravity compared to design errors as it can be cheaper to re-analyze the data or correct the reporting than redoing the whole study (Piantadosi, 1997). Meanwhile, published reports provide the main window for third parties to assess the quality of research including design and statistical analysis. For example reporting group means for paired data without reporting within pair changes may mislead the audience as to whether proper analyses or conclusions are made. Also in well conducted randomized trials, any difference in baseline characteristics between treatment groups can be attributed to chance and testing for statistical difference creates conceptional problems. This detailed analyses and reporting on testing equality of baseline characteristics between randomization groups is at the very least wastage of space.

CONCLUSION

The real solution to poor statistical reporting will come when authors learn more about research design and statistics; when statisticians improve their ability to communicate statistics to authors, editors, and readers; when researchers begin to involve statisticians at the beginning of research, not at its end; when more journals are able to screen more carefully more articles containing statistical analyses; and when readers learn more about how to interpret statistics and begin to expect, if not demand, adequate statistical reporting. For specialist statistical problems, see e.g. the paper by Chatfield (1991) and references therein. Thus we should be cautious about many potentially slippery patches as we like statistical excellence.

ACKNOWLEDGEMENT

I thank Anthony C. Ugwu for very helpful discussions and comments.

REFERENCES

Altman DG, Bland JM (1998). Generalization and extrapolation. BMJ 317(7155): 409-410.

Brett AS (1989). Treating hypercholesterolemia: How should practicising physician interpret the published data for patients? N. Engl. J. Med. 321: 676-80.

Creasy RK, Barrett CT, de swiet M, Kahanpaa KV, Kudolph AM (1972). Experimental intrauterine growth retardation in the sheep. AM. J. Obstet. Gynecol. 112: 566-573.

Dunnett CW (1970). Multiple comparisons. In: Mc Arthur JW, Colten T. eds. Statistics in Endocrinology. Cambridge, Mass: MIT Press; 79-103.

Ehrenberg AS (1981). The problems of numeracy. Am. Stat. 286: 67-71.

Evans M (1989). Presentation of manuscripts for publication in the British Journal of surgery . Br J. Surg. 76(131): 1-4.

Feinstein ARX, IPr P (1987). An improved summary for scientific communication J. Chromic. Dis. 40: 283-288.

Feinstein AR (1976). Clinical biostatistics XXXVII. Demeaned errors, confidence games, nonplussed minuses inefficient coefficients, and other statistical disruptions of Scientific Communication. Clin. Pharm. Therapeut. 20: 617-631.

Freiman JA, Chalmers TC, Smith H, Kuebler RR (1978). The importance of β the type II error and sample size in the design and interpretation of the randomized control trial. N. Engl. J. Med. 299: 690-694.

Gardner MJ, Altman D (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. BMJ. 292: 746-50.

Glantz SA (1980). Biostatisitics: How to detect, correct and prevent errors in the medical literature. Circulation, 61: 1-7.

Guyatt GH, Sackett DL, Cook DJ (1994). Users guide to the medical literature II. How to use an article about therapy or prevention. B. what were the results and will they help me I caring for my patients? JAMA 271: 59-63.

Hayran M (2002). Appropriate analysis and presentation of data is a must for good clinical practice. [My paper] Acta Neurochir. 83: 121-125.

Lang T, Secic M (1997). How to report statistics in medicine: Annotated guidelines for authors, editors, and reviewers. Philadephia (PA): American college of physicians.

Lu Y, Jin H, Chen MH, Gluer CC (2006). A procedure to evaluate odds ratios for osteoporotic fractures from different cross-sectional study cohorts. Osteoporosis int. 17(4): 507-520.

Mathews JNS (2000). An introduction to randomized controlled clinical trials. Londo: Arnold, pp. 37-49.

Murrary GD (1988). The task of a statistical referee Br. J. Surg. 75: 664-667.

Okeh UM, Ugwu AC (2008). Basic assumptions in statistical analyses of data in biomedical sciences. Int. J. Biol. Chem. Sci. 2: 1-2.

Piantadosi S (1997). Clinical trials. A methodologic perspective. New York: John Wiley & Sons, Inc; 62: 206.

Royston P, Altman DG, Sauerbrei W (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. Stat. Med. 25: 127-141.

Schwartz D, Lellouch J (1967). Explanatory and pragmatic attitudes in therapeutic trials. J. Chron. Dis. 20: 637-648.

Simon G, Wagner E, Vonkroff M (1995). Cost-effectiveness comparisons using “real world” randomized trials: The case of new antidepressant drugs. J. Clin. Epidemiol. 48: 363-373.

Thomas L, Krebs CL (1997). A review of statistical power analysis software. Bull. Ecol. Soc. Am. 78: 126-139.

Tukey JW (1977). Some thoughts on clinical trials, especially problems of multiplicity. Science. 198: 679-684.

US Geological Survey, Patuxent Wildlife Research Center. Web resources on power analysis; Available at:http//www.mp1-pwrc.usgs.gov/powcase/powlinks.html.

Wallenstein S, Zucker CL, Fleiss JL (1980). Some statistical methods useful in circulation research. Circ. Res. 47: 1-9.

Young JL (1999). Biostatistics and Clinical trials. A view J. stat. Plan. Inference. 79: 349-367.