Statistical Pitfalls in Medical Research

VB Nyirongo[1], MM Mukaka[1], LV Kalilani-Phiri[2]

1 MLW Clinical Research Programme, College of Medicine, Malawi.
2 Research Support Centre, College of Medicine, Malawi.

Correspondence: Dr. VB Nyirongo, VB, P.O. Box 30096, Chichiri, BT3. Email: vnyirongo@mlw.medcol.mw

Abstract

In conducting and reporting of medical research, there are some common pitfalls in using statistical methodology which may result in invalid inferences being made. This paper is aimed to highlight to inexperienced statisticians or non-statistician some of the common statistical pitfalls encountered when using statistics to interpret data in medical research. We also comment on good practices to avoid these pitfalls.

Introduction

Statistical methods in medical research allow inferences to be extended beyond study subjects to the study population e.g. future patients[1]. 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[2]. However there are many pitfalls that are encountered when using these statistical methods. Statistics is a huge discipline with different paradigms, schools of thought and alternative methodologies such that sometimes rationales for choosing one method over the other can be confusing. Inappropriate statistical methods are sometimes used when designing a study, selecting the types of variables, the distributions of variables, the number of groups compared or the dependency structure among these groups.

The primary goal of the analysis and the study design determine the appropriate statistical analysis method to be used[3]. It is also easy to overlook statistical and mathematical assumptions of different methods. In this paper we discuss some of the common statistical landmines during study design, analysis and interpretation of the results. This is primarily intended for inexperienced statisticians or non-statisticians.

Study design pitfalls

It is important to choose the right study design in order to answer the research question in a cost effective manner. The study design may influence the study 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. It should be stated that not seeking statistical advice on study design is one of the commonly encountered statistical "hazards" in research.

Apart from choosing the most effective study design and sample size calculations, this stage also involves specification of main hypotheses, outcomes, potential confounding or risk factors; and for randomized controlled trials, defining randomization and blinding procedures.

We highlight some of the common errors encountered involving sampling plan, sample size calculations, laboratory assays and randomizations.

Sampling plan

Proper sampling is a precondition for avoiding bias. For example in assessing prevalence of HIV, a sample of special groups e.g. pregnant women cannot represent the general population as pregnant women are highly sexually active individuals. Another important pitfall related to the sampling exercise in medical research is non-response e.g. in behavioural health studies, non-response can easily be due to self-selection which introduces selection bias.

Sample size calculations

It is ideal that the sample size is calculated to obtain estimates of desired precision or to detect an effect if it exists (e.g. a minimum detectable difference between two treatments). A smaller sample than required would not have enough power for statistical conclusions. Obviously, unnecessarily larger samples would require more resources than could be justified by the gain in precision or power to detect the difference. The following points should be seriously considered when calculating a sample size for a study:

1. Sampling technique used e.g. simple random sampling or cluster sampling.
2. Variability in the population i.e. how individual data points vary around the expected value.
3. Accuracy of the estimate or detectable difference required and
4. The statistical model or test e.g. paired t-test or two independent sample t-test to be used for analysis.

It is common to have drop-outs or loss to follow-up in a cohort study. This is especially common in populations that are transient, such as migrant groups or labourers who shift from one geographic region to the next in search of employment. If the number of study participants who are lost to follow-up is large, it can 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. Therefore, when calculating sample size, it is important 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.

Laboratory assays

When conducting a study involving 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. A common mistake in sample pooling is not considering the probability of samples testing positive, and calculating the expected number of tests to be done, which may result in testing 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.

Randomisation

In experimental clinical studies, the primary aim 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 the groups are as balanced in terms of all other factors (both known and unknown) as possible. It is easy to adjust for known potential confounders at analysis stage, but not the unknown factors. Randomisation is one of the statistical tools used to ensure that treatment groups are balanced. If randomisation is done correctly, any imbalances between groups are due to chance alone. Randomisation using blocks ensures that the numbers of participants are balanced between groups. Blocking is particularly necessary in small studies because simple randomisation can lead to imbalance in the number of participants in the trial arms, which could reduce the power of a study[4, 5]. However caution 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 randomisation used include stratification and minimisation techniques to ensure balancing with respect to known prognostic factors[4, 5].

We should highlight that errors at the study design stage usually have high gravity stemming from the fact that design errors cannot be corrected once the study has been done[4]. In the next section we highlight some of the common errors encountered involving statistical analysis.

Statistical analysis and reporting pitfalls

It is ideal to consider for analysis only hypotheses, outcomes, potential confounding or risk factors pre-specified during study design so that strong inferences can be made from the results. Pre-specifying main hypotheses avoids the problem of multiple testing and data snooping/dredging during analysis[5].

Subgroup analysis

Ad hoc subgroup analyses are vulnerable to data dredging. Ideally results of such analysis should be viewed as exploratory[6]. Even with subgroups specified before seeing the data, subgroup analyses introduce multiple testing which increases the chance of obtaining a false positive result and therefore should be corrected for[7]. On the flip side, there is a high chance of missing a true treatment effect due to small sample sizes in subgroups. Also subgroup analyses cannot be used to assess interaction between factors (interaction between two variables exists when there is a difference in effect of one variable on the outcome across another variable). For example, it is wrong to conclude that there is interaction by looking at two subgroups and finding a significant difference in one but not in the other[6,7,8].

A special case of subgroup analysis is removing seemingly outlying observations. Outliers might represent an important aspect of the system and careful consideration including sensitivity analysis should be done before removing them.

Effect measure modification/interaction

Effect measure modification/interaction refers to the extent to which the joint effect of two risk factors on disease differs from the independent effects of each of the factors[8,9]. One example is phenylketonuria (PKU), a metabolic disorder in which the combination of a genetic mutation and an environmental factor, in this case dietary exposure to a particular amino acid gives rise to mental retardation in children[10]. Statistical interaction is a model-dependent concept and sometimes a chosen model cannot represent the underlying biological mechanisms of causality[10, 11, 12, 13]. Statistical interaction depends on the form of statistical model used to estimate a measure of effect e.g. a risk measure (risk difference, risk ratio or odds ratios). Statistical interaction can also depend on the scale of analysis. It is possible to find an interaction on an additive scale and no interaction on the multiplicative scale. However, biological interaction refers to the interaction of two factors in causing disease. It is therefore important for investigators to clearly define the interaction being assessed and use correct parameters when assessing the interaction.

Confounding

It is not always possible to conduct randomized controlled trials (RCTs) which are considered the gold standard because of ethical, economical and other reasons. Alternatively, observational studies are used to explore and potentially infer causality. However, unlike RCTs, there is usually an imbalance of prognostic factors in different exposure groups leading to confounding (i.e. masking the effect of the exposure variable on the outcome) and bias. A potential confounder is a variable that is associated with the exposure variable and also influences the outcome[14, 15]. A classic example of confounding is the initial association between alcohol consumption and lung cancer which is confounded by smoking (smoking is associated with alcohol use, an independent risk factor for lung cancer).

Also in Genome Wide Association studies, recently becoming increasingly popular, confounding or spurious associations due to population admixture are a major concern. A simple example is a study with cases and controls coming from different ethnic groups with different allele frequencies for a particular genedue to different ancestry.

Common methods used to account for confounding in the analysis stage are stratification or including potential confounders in a regression model. Statistical analysis can only control for confounding that has been measured, and does not control for unmeasured confounders. Additionally, adjustment only takes into account the extent to which the confounders have been measured. In some studies the measurement of confounders is not as rigorous as for the main exposure and outcome variables. Inadequate measurement of confounders can introduce bias[11,12]. Caution

should also be taken not to adjust for variables that are in the causal pathway as this may create spurious associations and in some cases attenuate, inflate or remove a true effect[16, 17].

Model choice

Choosing aspects of the study design to model is very important. For example, ignoring some characteristics like dependence among observations can result in inefficient estimators[18]. Dependence occurs when data is 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-test for paired data is clearly inappropriate.

Model choice also encompasses 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 information.

Another aspect of choosing a model involves selecting explanatory variables to enter the model. When selecting the variables to enter a model, one needs to be aware that putting a lot of variables into the model may result in unstable estimates especially when a set of explanatory variables are highly correlated with the exposure[19]. Also inclusion of variables unrelated to the outcome increases unexplained variability. On the other hand selecting a few variables into the model may also result in biased estimates if the omitted variables are true confounders.

Furthermore, expert knowledge might require including variables in the model which otherwise would not have been chosen using mathematical criteria alone. Also note that different mathematical criteria or automated procedures can lead to different sets of variables being selected.

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 categorisation should be done only when necessary[20, 21, 22].

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 e.g. 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 have missing data due to non-response.

Another form of missing data is loss to follow-up e.g. in a study of HIV infected individuals where the outcome is morbidity or mortality, patients maybe lost to follow-up if they were too sick to come for follow-up visits or died, and the researchers were 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.

Interpretation of results

It is common to report only p-values for statistical tests. However, confidence intervals for estimates are more informative than p-values. A p-value depends on both the size of the difference in the groups and precision (i.e. sample size and population variability) [8, 23]. Large studies with more precision may give small p-values even if the difference is not clinically important[20]. On the other hand small studies with less precision but a large difference between groups will also give a small p-value. Therefore a p-value does not tell us whether the significance is due to effect size or sample size. However from the confidence interval both precision of the estimate and observed difference between groups are available.

It is also a common mistake to perceive a significant statistical association as sufficient basis for inference on causation. Statistical association is only one of the required factors. More information e.g. biological plausibility is needed to declare causation.

Statistical software

Statistical software 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.

Reporting pitfalls

Arguably, errors conducted during analysis or reporting stage usually have relatively low gravity compared to design errors as it can be cheaper to re-analyse the data or correct the reporting than redoing the whole study[4]. However these mistakes are no less important as 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. Thus detailed analyses and reporting on testing equality of baseline characteristics between randomisation groups is at the very least wastage of space.

Using graphical tools

Figures and tables should not be used to "store" data! [20] 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 summarised data and information in a useful and non-misleading manner and to aid interpretation of the results.

Discussion

We have highlighted some common statistical dangers in medical research involving design, analysis and reporting. Gross mistakes can be minimised by distinguishing exploratory from confirmatory analyses[24]. Statistical analyses in experimental studies (randomized clinical trials) should be limited to those pre-specified in the study protocol. On the other hand exploratory analyses in observational research require deciding a priori on the model type before parameter estimation. Thus aims of a study are very important and should be stated clearly at the beginning. Controlling for confounders by randomisation rather than adjusting for them in the regression model can avoid having highly linearly correlated explanatory variables in the model (multicollinearity problem) i.e. confounders (both known and unknown) are controlled at the design stage4, [25]. Randomisation is the only method that can balance unknown confounders between treatment groups. Controlling for confounders by randomisation is possible when interested in hypothesis testing of marginal associations only. To calculate adjusted estimates or conditional associations, known and measured confounders must be controlled in analysis as well. Some methods of controlling for confounders in the design stage e.g. matching do not permit estimation of the effect of the confounder during analysis. On the other hand the analysis for effects always has to take into account the effect modifiers and confounders if marginal effects are different to the effects in subgroups. This should include taking into account effect modifications during sample size calculation, in order to have enough power to detect them if they exist.

Aims of the study can help in deciding whether to do subgroup analyses or not. Doing subgroup analysis when effects estimation is the aim and subgroups and marginal effects are similar (for example there is homogeneity of odds ratio across subgroups) unnecessarily reduces the effective sample size resulting in reduced precision for the estimates[4], [26].

Additionally, study objectives, design and power of the study should always be considered in interpreting the results. There should be a distinction between "pragmatic" (effectiveness) and "explanatory" (efficacy) studies when designing and interpreting biomedical research. When the aim of the analysis is to describe or explain the phenomenon, estimates should be adjusted for prognostic/risk factors[5]. On the other hand if estimation of the effects is the aim of analysis, confounders, including effect modifiers have to be included in the model. For hypothesis testing, analysis should adjust for all baseline characteristics that are potential risk factors and were specified at study design. P-values have also to be adjusted to account for multiple testing in subgroup analyses, ideally specified at study design. Otherwise any ad hoc subgroup analysis would be exploratory and no strong inference can be made lest one be accused of data dredging[27].

In analysis, assumptions of statistical methods or models used should be fulfilled satisfactorily or checked e.g. linear regression analysis should only be done after establishing that the relationship between the response and explanatory variables is linear and the variability of responses is normal with constant variance. A common assumption in statistics is the independence of observations; however dependent data are very common in practice. Ignoring the dependency structure between observations results in either under-estimating standard errors, or inefficient estimators, depending on the "within" and "between" variability for groups of dependent observations.

This discussion paper has highlighted a few statistical pitfalls and our list is by no means exhaustive. For specialist statistical pitfalls see e.g. the paper by Chatfield[28] and references therein. Thus we should be cautious about many potentially slippery patches as we hike to statistical excellence.

Acknowledgement

We thank Terrie Taylor and Sarah White for very helpful discussions and comments.

References

1. Altman DG, Bland JM. Generalisation and extrapolation. BMJ 1998; 317(7155):409-410.
2. Young JL. Biostatistics and clinical trials: a view. J. Stat. Plan. Inference 1999; 78: 349-367.
3. Hayran M. Appropriate analysis and presentation of data is a must for good clinical practice. [My paper] Acta Neurochir Suppl. 2002; 83:121-125
4. Piantadosi S. Clinical trials: a methodologic perspective. New York: John Wiley & Sons, Inc; 1997: 62, 206.
5. Mathews JNS. An introduction to randomized controlled clinical trials. London: Arnold;2000:37-49.
6. Altman, DG, Matthews JNS. Interaction 1: Heterogeneity of effects. BMJ 1996; 313(7055): 486.
7. Cook DI, Gebski VJ, Keech AC. Subgroup analysis in clinical trials. Med J Aust 2004; 180(6):289-291.
8. Matthews JNS, Altman, DG. Interaction 2: Compare effect sizes not P values. BMJ 1996; 313(7060): 808.
9. Miettinen O. Confounding and effect-modification. Am J Epidemiol 1974; Nov; 100(5):350-3.
10. Ahlbom A, Alfredsson L. Interaction: A word with two meanings creates confusion. Eur J Epidemiol 2005; 20(7):563-4.
11. Rothman KJ and Greenland S (editors) (1998) Modern epidemiology. Philadelphia: Lippincott-Raven; 1998: 329-342
12. Rothman KJ (2002) Epidemiology: An Introduction. Oxford: Oxford University Press; 2002.
13. Greenland SG, Rothman KJ. Concepts of interaction. In: Rothman KJ, Greenland S (eds). Modern Epidemiology. Philadelphia: Lippincott-Raven Publishers; 1998:329–342.
14. Greenland S. Quantifying biases in causal models: classical confounding vs. collider-stratification bias. Epidemiology 2003; 14:300–6.
15. Hernán MA, Hernández-Diaz S, Werler MM and Mitchell AA. Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. Am. J. Epidemiology 2002; 155:176-184
16. Cole SR, Hernan MA. Fallibility is estimating direct effects. Int J Epidemiol 2002; 31:163–5.
17. Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. Int J Epidemiol 1980; 9:361–7.
18. Poirier DJ and Ruud PA. Probit with dependent observations. StudentBMJ 2003; 11:349-392. The Review of Economic Studies 1988; 55: 593-614.
19. Greenland S. Modeling and variable selection in epidemiologic analysis. Am J Public Health 1989; 79(3):340-9.
20. Lang T. Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles. Croatian Medical Journal 2004; 45(4):361-370.
21. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/CatContinuous.
22. Royston, P., Altman, D. G., & Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: a bad idea. Statistics in Medicine 2006, 25,127-141.
23. Royall RM. The Effect of Sample Size on the Meaning of the Significance Tests. The American Statistician 1986; 40(6): 313-315.
24. Altman DG. Practical statistics for medical research. London: Chapman & Hall; 1991:338,465,466.
25. Hennekens CH, Buring JE and Mayrent SL (ed). Epidemiology in Medicine. Boston: Little, Brown, 1987.
26. Assman SF, Pocock SJ, Enos LE and Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet 2000; 355:1064-1069.
27. Gebski VJ and Keech AC. Statistical methods in clinical trials. The Medical Journal of Australia 2003; 178 (4): 182-184.
28. Chatfield C. Avoiding Statistical Pitfalls. Statistical Science 1991; 6: 240-252.