RESEARCH SERIES

Statistical consideration for research

S Carley, F Lecky

Emerg Med J 2003;**20**:258–262

The seventh paper in this series discusses the importance of statistical techniques in research.

The lack of statistical support and knowledge has been cited as one of the most important obstacles to research in emergency medicine.[1] This is perhaps not surprising as for many clinicians statistics is seen as an academic exercise necessary only to get a piece of research published in a journal. Others may find the close relation of statistics to the mathematics learnt many years previously so far removed from their current knowledge that it seems incomprehensible.

Unfortunately, the relevance of statistics to research and medicine is often taught at such an early stage in a medical career that it often seems irrelevant until a personal interest in research develops later. Yet, while it is true that statistics are an important component of research, the requirement for authors of scientific papers to show some statistical interpretation of their results is sometimes not reflected in the appropriate use of statistical techniques.[2]

This article seeks to show why it is important for the researcher to have some understanding of statistical techniques. However, an article of this size cannot attempt to summarise the huge topic of statistics, and in fact this has been done extremely well in the series by Driscoll *et al* published in this journal.[3] Rather, it is our aim to show that an appreciation of statistical analysis, with appropriate support, is essential for performing effective research. Some indications will be given as to which statistical tests are best suited to the various types of study, and of the limitations of statistical inference.

WHY IS STATISTICAL ANALYSIS IMPORTANT?

While the answer to this question may seem obvious it is important to understand the reasoning behind statistical analysis and why it is important to use mathematical tests on data to assess whether they support or refute study hypotheses.

The vast majority of medical research is conducted with a sample of subjects (for example, patients) taken from a wider population. Populations are large groups of people in a defined setting (for example Manchester emergency department attenders) or with a certain characteristic (for example, Scaphoid fracture). It is rarely possible to study the whole of a population, but it is possible to take a subset of a population to study. A subset of a population is called a sample. Regardless of the type of study conducted, at the end of the project the researcher is left with a set of data based solely on the sample examined. However, with all observations regarding injury or disease there is the possibility that the measurements made on the sample may misrepresent the population as a whole because of chance. When looking at a sample of patients it is unlikely that the sample will behave in precisely the same way as if it were possible to enrol the whole population in the study. The difference in the behaviour of a sample from the true population because of chance alone is known as random variation. Generally the smaller the number of patients in a sample the greater the likelihood that the result reflects random variation rather than the true population result.

For example, an emergency department based study looked at two techniques for reducing Colles fractures in the emergency department.[4] The population for this study could be regarded as all patients with Colles fractures requiring manipulation presenting to every emergency department. Clearly it would be impossible, (and indeed unnecessary), to study all the patients in this population. However, a sample of patients presenting to a single (or several) emergency departments with Colles fractures could be entered into a clinical trial to compare the two different interventions. Box 1 illustrates the study by Kendall *et al*.

Table 1 shows that in both groups patients experienced some pain. By examining the table we can see that median pain scores were less after a Biers block at administration and manipulation but slightly higher after 30 minutes. However, pain is multifactorial and may have been influenced both by the technique and by the characteristics of the patient themselves. The difference may be attributable to a real difference between the two techniques or to random variation.

We can use statistical analysis to estimate how likely it is that the results may have arisen purely by random variation (chance). In this study, a

Box 1

One hundred and forty two patients with Colles fractures were randomised to having a reduction performed under haematoma block or Biers block anaesthesia. Principal outcome measures included pain and remanipulation rate. Pain was assessed using a 10 point visual analogue scale. The null hypothesis for the study was that there would be no difference in pain scores or manipulation rates between the two techniques (a hypothesis was not given by the authors in the published study). Table 1 shows the principal results.

See end of article for authors' affiliations

Correspondence to: Mr S D Carley, Department of Emergency Medicine, Manchester Royal Infirmary, Oxford Road, Manchester, UK; s.carley@btinternet.com

Emerg Med J: first published as 10.1136/emj.20.3.258 on 1 May 2003. Downloaded from http://emj.bmj.com/ on October 2, 2022 by guest. Protected by copyright.

Table 1

	Median pain scores			
	Pain on administration of anaesthetic	Pain at manipulation	Pain 30 minutes post manipulation	Remanipulation rate
Haematoma block. n=72	5.3	3.0	1.0	24%
Biers block n=70	2.8	1.5	1.2	6%
Statistical test	Mann-Whitney U	Mann-Whitney U	Mann-Whitney U	χ^2
p Value	<0.001	<0.01	Not significant	0.003

Adapted from: Kendall JM, Allen P, Younge P, *et al*. Haematoma block or Biers block for Colles' fracture reduction in the accident and emergency department—which is best?. *J Accid Emerg Med* 1997;**14**:352–6.

Key point

- To apply appropriate statistical tests, enough good quality data must have been collected.

Mann-Whitney U test was used to test the pain scores. This showed that the difference in pain at administration and during manipulation was very unlikely to have arisen by chance though the difference at 30 minutes may be attributable to random variation. Examination of the remanipulation result using a χ^2 test showed that the higher remanipulation rate with a haematoma block was unlikely to have arisen by chance. If the statistical analysis shows that the differences found are unlikely to have arisen by chance then we accept that they are attributable to the difference in anaesthetic techniques. We can therefore conclude that Biers block is a better technique in terms of analgesia and remanipulation rate.

Statistical tests, particularly if presented with confidence intervals help estimate how likely chance and random variation could have had a bearing on the study result. Although it is impossible to eliminate all uncertainty in the results of a study, the assessment of how likely the results may have arisen by chance is an important factor when deciding whether or not the results are convincing enough to subsequently influence clinical practice. In fact statistical analysis alone should never influence clinical practice as explained below. In contrast with popular opinion then, it is impossible to "prove" any result with statistics; rather good statistical analysis tells us the probability that the results in the study reflect the situation in the population as a whole.

WHEN TO OBTAIN STATISTICAL HELP

It is a common misconception that the best time to approach a statistician for help is when analysing the data from an already completed study. In fact there is perhaps no better way to frustrate a statistician than to adopt this approach. Seeking statistical help once the study has been completed implies that the purpose of statistics, and the contribution from statisticians, is solely in support of the analysis of data. Considering statistical issues at a late stage in a research project makes it impossible for past mistakes in the project methodology to be identified and corrected. In fact it has been suggested that research design is arguably the most important aspect of the statistical contribution to medicine.[3]

Clearly, statistics is rooted in the analysis of data, no clever analysis can compensate for problems with the design of a study. It is therefore essential that consideration of statistical issues take place at all points of a study including the design stage. A study can be thought of as consisting of several stages (fig 1). Many would agree that data processing and analysis falls within the remit of statistics but this cannot be competently achieved without good design and planning. Analysis must therefore be considered from the earliest stages of a study.

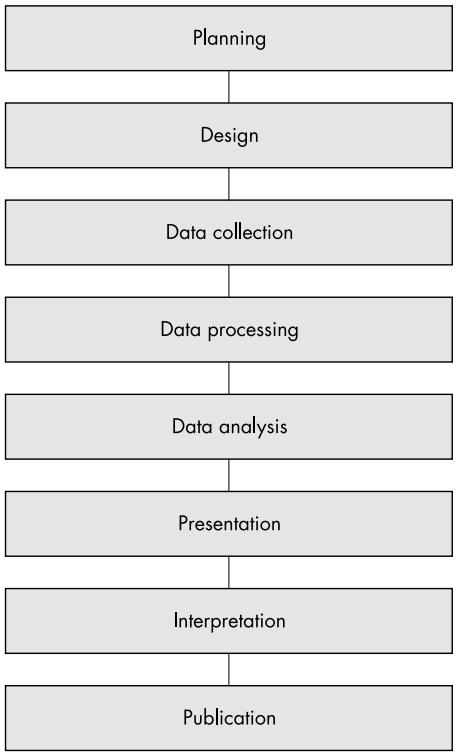

Figure 1 The eight steps in completing a research project.[4]

PLANNING A STUDY

A formal hypothesis or set of objectives for the study should be derived during the planning stage. This then facilitates clear identification of the best study design.

STUDY DESIGN

It is at this stage that the researcher (along with a statistician) should select the statistical test that will be used to analyse the data. The type of test will be dependent on the study hypothesis. Table 2 gives some guidance.

Once the appropriate statistical measures and methods of comparison have been identified the numbers of patients required can be predicted from a power study.

POWER AND SAMPLE SIZE CONSIDERATIONS

Studies that compare the outcomes between two or more groups may come to one of two conclusions. Either that there is no difference between the two groups or that a difference does exist. We have seen how statistical significance can help the researcher assess how likely each of the outcomes are. However, when a result is obtained there are two important errors that may occur. Firstly, it is possible that an analysis of the sample data demonstrates a difference when one does not

Emerg Med J: first published as 10.1136/emj.20.3.258 on 1 May 2003. Downloaded from http://emj.bmj.com/ on October 2, 2022 by guest. Protected by copyright.

Table 2 Common statistical tests used in hypothesis testing

Hypothesis relates to	Potentially suitable statistical tests/measures
Relation between characteristics	Regression analyses, other models
Risk factors for disease in populations	Odds ratios, risk ratios
Performance of diagnostic strategies	ROC curves, comparisons of sensitivity, specificity, likelihood ratios
Effectiveness of treatments	Relative risk ratios, survival analyses
Differences between groups	*t* tests, χ^2 analyses, analysis of variance

Key point

- A researcher should obtain statistical help before embarking on a study to ensure that an adequate number of patients are enrolled in the trial.

truly exist in the population. This is known as a type 1 statistical error (a false positive result). The second type of error occurs when analysis fails to demonstrate a difference between the samples when there really is one in the population. Failing to demonstrate a difference when one truly exists is known as a type 2 statistical error (a false negative result).

The ability of a research project to avoid making a type 1 or 2 statistical error is referred to as the power of a study. It is an aspect of research that is often forgotten or ignored.[6] As the power of a study is of such importance when interpreting the results of a study it should come as no surprise that it is also a key aspect of research design. Despite this, many papers in the medical literature fail to demonstrate any consideration of power either before or after a study has been conducted. In many cases sample sizes seem to have been based on time[7] or convenience[8] rather than the number needed to be assured of showing a true difference between interventions.

The possibility of making a type 1 or 2 error through lack of power depends on a number of factors.

1 Sampling

The sample of patients taken from the population of interest is subject to a degree of random variation. It is therefore possible to pick a sample that does not represent the population. However, the likelihood of selecting a sample that significantly differs from the population decreases if a larger sample size is used. Therefore, the results from looking at a large sample of patients are less likely to be incorrect than if only a small number of subjects are examined.

2 Event rate

The event rate being examined also influences the likelihood of getting a false result. A rare event requires a large number of patients to be entered in a trial in order to recruit enough patients with the event in question to demonstrate a statistically significant result. For example, we may wish to conduct a five year study to reduce suicide rates among men aged 20–30 attending emergency departments. Such a study would require an enormous number of patients, as comparatively few men aged 20–30 will commit suicide. However, if we were to conduct the study using alcoholic men presenting to the emergency department after deliberate self harm a smaller number of patients would be needed, as more of these patients are likely to commit suicide in the next five years.

3 Random variation

When looking at a study that produces data showing random variation, for example blood pressure, there is a degree of variability between subjects. This can be expressed with summary statistics such as standard deviation. This variability of the outcome in question also affects the ability of a study to give a true result. In studies comparing outcomes with a wide variability a large number of subjects is needed in order to detect a clear difference in outcome.

Box 2

A study is planned to investigate the impact of a new type of paramedic with extended skills for the management of out of hospital VF arrest. We know from previous studies that the survival rate is in the region of 10% at present.[9]
To estimate the number of patients needed in such a study we need to specify the following.

- 1 We are interested in finding out if the new paramedics can increase survival by 5%.
- 2 We will perform a χ^2 analysis on the data
- 3 As the outcome measured is live or dead at hospital discharge then issues of variability do not concern us.
- 4 We are willing to accept a 5% possibility of getting a type 1 error (finding a difference if there isn't one)
- 5 We want to be 80% certain of detecting a difference of 5% or more if it truly exists (that is, 80% certain of not making a type 2 error).

Using a nomogram method[5 10] we can calculate that the number of patients required would be in the region of 3500 (or 1750 in each group). To calculate the exact number we could use one of a number of statistical computer programs.

4 Statistical factors

The type of data collected influences the choice of statistical test that will be used (table 2). Power calculations are also affected by the way in which the data are to be analysed. It is therefore vital that the researcher and statistician meet at an early stage in the design of a trial to discuss what type of data are to be collected and how the researcher wishes to interpret the data. Only then can an estimate of the number of patients required in the trial be made.

Many of the formulas used in power calculations are not straightforward and we advise the help of a statistician for all but the most basic studies. However, box 2 shows the type of information and results found in a hypothetical study based on data from out of hospital cardiac arrest.[9] For simple clinical trials nomograms or graphs can be used to estimate the sample size required for the study.[10] Although graphical methods work well they often make several assumptions about the data being analysed, and about the type of statistical tests to be used. Consequently, it is all too easy to incorrectly estimate the power of a study without a good working knowledge of statistics. If in any doubt, we would recommend that formal statistical help be obtained to calculate the power of a study before embarking on a project.

Calculating the power of a study at the design stage is increasingly a requirement when seeking ethical committee approval for a research project. It is an ethical issue as it is unfair to subject patients to an experiment that is too small to produce a meaningful result (producing a type 2 error). Similarly, a trial should not recruit a greater number of patients

Emerg Med J: first published as 10.1136/emj.20.3.258 on 1 May 2003. Downloaded from http://emj.bmj.com/ on October 2, 2022 by guest. Protected by copyright.

than is necessary to answer the original question. This would be unethical from the patients perspective and wasteful of resources. Unsurprisingly, it is rare for the second of these two scenarios to occur in practice!

Performing a power calculation before starting a study also avoids the temptation to reanalyse results after every patient is recruited into a trial up until a point is reached at which the results become statistically significant. If results are sequentially analysed until a statistically significant effect is reached the possibility of declaring a positive result when one does not truly exist (a type 1 error) may be as high as 30%. This approach must therefore be strongly discouraged.

ANALYSIS AND REPORTING OF RESULTS

An important aspect of statistical method is the clear numerical and graphical presentation of results.[9] For many research projects the publication or presentation of results in a peer reviewed journal is an important concluding part of a project. It is therefore important that the results of the study are presented clearly. However, errors in the presentation of results are common in the medical literature. Data can be presented in numerical or graphical form and there are pros and cons with both approaches. Some data can only be presented in one form but many results can be presented as either. Unfortunately many medical journals do not allow authors to duplicate results that appear in the text as tables or graphs and some judgement must be made as to the best way to present the findings. Common errors in the presentation of results are explained in detail elsewhere[11] but many mistakes are made because of a basic lack of understanding of the data. The finding of a result that is statistically significant or one that can be expressed as a summary statistic should not discourage the author from presenting additional information. Papers may even reach publication with results that appear as statistically significant ($p<0.05$) with no mention of the method used to obtain these figures.[12] This may lead to scepticism regarding the analysis on the part of the reader and question the information presented. As most of the errors in presentation are the result of a misunderstanding of the data, it may be advisable to discuss the presentation of the results when seeking statistical advice on the analysis of the study data.

Part of the problem may be attributable to the widespread availability of extremely powerful statistical packages for personal computers. This can be viewed as both a blessing and a curse as it is now comparatively easy for the computer literate researcher to perform a large number of statistical analyses on their data without ever really understanding the statistical process.

In particular, there is a concern at the use of the "black box" technique to analyse data. This is when a computer is used to perform an incorrect statistical method because there has been no understanding of the basic data, or when a large number of statistical tests are performed in pursuit of a "statistically significant" finding. Drawing conclusions from an incorrect analysis is simply misleading whereas searching for a statistically significant result to influence publication is unethical. The medical journals peer review process should identify poor statistical methods and black box analysis but it still occasionally passes critical appraisal and reaches publication.[13]

CLINICAL AND STATISTICAL SIGNIFICANCE

The difference between results that are statistically significant and results that are clinically important is vital in understanding the contribution of statistical analysis to research. Statistical analysis tells us how likely the findings of a study are to have arisen by chance. The term "statistical significance" is often used to denote a result that is important and correct yet this is a misinterpretation of what statistical analysis tells us about the data. Most journals now encourage authors' to present results of studies together with a measure of the statistical significance attached to the results. This is usually presented as a "p" number with p being shorthand for probability.

Results attached to a $p<0.05$ tag are commonly regarded as being statistically significant. In this regard the term *statistically significant* is misleading, as it is merely an indication of the plausibility of the null hypothesis. It represents the probability of finding the results if the null hypothesis in question is correct (that is, that there really is no difference between the groups studied). Conventionally the significance level of 5% ($p=0.05$) is considered as statistically significant. This means that if there really were no difference between the groups studied then a result at least as unlikely as the one found would occur on 5% of occasions. The 5% level is an arbitrary one and, in our opinion, represents a fairly high level of uncertainty on which to base clinical management. Significance levels therefore can only be used as a guide to the interpretation of the data. Other levels of significance that are commonly quoted are 1% ($p=0.01$) and 0.1% ($p=0.001$), obviously these results are much less likely to have arisen by chance.

A further difficulty occurs when examining the results of trials with multiple analyses. The more analyses performed, the more likely it is that one of the analyses will turn up a result that is *statistically significant*. If a study contains 20 different analyses of data then it is likely that one of the results will be turn out to be significant at the 5% level by chance alone.

The finding of a "statistically significant" result does not mean that the results are important or that they should change clinical practice. Clinical practice should also be based on the difference to the outcome of the patient. In fact, clinical significance has been defined as an unbiased finding that changes clinical practice.[14] In its broader sense the term implies that the findings are not only unlikely to have arisen by chance, but that they are also important in clinical care.

For example, in a study comparing the antipyretic efficacy of ibuprofen and paracetamol in children with febrile seizures, the authors found that at four hours after administration the temperature in both groups of patients had fallen but that those patients receiving ibuprofen had an average temperature 0.5°C less than those receiving paracetamol.[15] This was described as statistically significant ($p=0.05$). However, the difference in temperature occurred only at the four hour reading and was not sustained beyond this time. Its clinical importance is therefore probably unimportant. The distinction between statistical and clinical significance is important for the researcher, as it is important to seek results that are clinically important, rather than those that satisfy statistical tests.

The search for statistically significant findings rather than clinically important ones would be less likely if there was not an inherent publication bias for articles showing positive (statistically significant) results.

CONCLUSION

It is a misconception to see statistical analysis as a means to an end, this being getting a paper accepted for publication. In fact statistical considerations encompass all aspects of the research process. It is therefore foolish for the emergency medicine researcher to only consider statistical matters once a set of data has been collected at the end of a project. Identifying the right statistical test is an essential component of research design and this should ensure that data are obtained correctly with sufficient power to address the original research question.

Even the most complicated statistical analysis can never make up for a research design that is flawed. It is therefore vital to ensure that the design of any research is as watertight as possible. If in any doubt, researchers may benefit greatly

Emerg Med J: first published as 10.1136/emj.20.3.258 on 1 May 2003. Downloaded from http://emj.bmj.com/ on October 2, 2022 by guest. Protected by copyright.

from involving expert statistical help at all stages of a research project. For all except the most simple of studies expert statistical advice should be sought when considering sample size, data collection, data analysis, and presentation for publication.

Authors' affiliations
S Carley, Department of Emergency Medicine, Manchester Royal Infirmary, Manchester, UK
F Lecky, Department of Emergency Medicine, Wythenswawe Hospital, Manchester, UK

REFERENCES

1 **Cooke M**, Wilson S. Obstacles to research in A&E. *J Accid Emerg Med* 1997;**14**:269.
2 **Gore SM**. Misuse of statistical methods: critical assessment of articles published in the BMJ from January 1976 to March 1976. *BMJ* 1977;**1**:85–7.
3 **Driscoll P**, Lecky F, Crosby M. An introduction to everyday statistics—1. *Emerg Med J* 2000;**17**:205–11.
4 **Kendall JM**, Allen P, Younge P, *et al*. Haematoma block or Biers block for Colles' fracture reduction in the accident and emergency department—which is best? *J Accid Emerg Med* 1997;**14**:352–6.
5 **Altman DG**. *Practical statistics for the medical researcher*. London: Chapman and Hall, 1992.
6 **Becker PJ**,Viljoen E, Wolmarans L, *et al*. An assessment of the statistical procedures used in original papers published in the SAMJ during 1992. *S Afr Med J* 1995;**85**:881–4.
7 **Howell MA**, Guly HR. A comparison of glucagon and glucose in prehospital hypoglycaemia. *J Accid Emerg Med* 1997;**14**:30–2.
8 **London NJ**, Osman FA, Ramagopal K, *et al*. Hyaluronidase (Hyalse): a useful addition in haematoma block? *J Accid Emerg Med* 1996;**13**:337–8.
9 **Gore SM**, Altaman DG. Presentation of results. In: *Statistics in practice*. London: BMJ Publishing Group, 1982.
10 **Gore SM**, Altman DG. How large a sample. In: *Statistics in practice*. London: BMJ Publishing Group, 2001:6–8.
11 **Altman DG**. Statistics in published papers. In: Altman DG, ed. *Practical statistics for the medical researcher*. London: Chapman and Hall, 1992:481–97.
12 **Mann CJ**, Heyworth J. Comparison of cardiopulmonary techniques using video camera recordings. *J Accid Emerg Med* 1997;**13**:198–9.
13 **Crombie IK**. *The pocket guide to critical appraisal*. London: BMJ Publishing Group, 1996.
14 **Sackett DL**, Haynes RB, Guyatt GH, *et al*. *Clinical epidemiology: a basic science for clinical medicine*. Boston: Little, Brown, 1991.
15 **Van Esch A**, Van Steensel-Moll HA, Steyerberg EW, *et al*. Antipyretic efficacy of ibuprofen and acetominophen in children with febrile seizures. *Arch Pediatr Adolesc Med* 1995;**149**:632–7.

Emerg Med J: first published as 10.1136/emj.20.3.258 on 1 May 2003. Downloaded from http://emj.bmj.com/ on October 2, 2022 by guest. Protected by copyright.