1 Uses and abuses of medical statistics

1.1 Introduction 2
1.2 Why use statistics? 3
1.3 Statistics is about common sense and good design 4
1.4 Types of data 5
1.5 How a statistician can help 9
1.6 Further reading 11
1.7 Exercises 12

Medical Statistics Fourth Edition, David Machin, Michael J Campbell, Stephen J Walters

Summary

Statistical analysis features in the majority of papers published in health care journals. Most health care practitioners will need a basic understanding of statistical principles, but not necessarily full details of statistical techniques. Medical statistics can contribute to good research by improving the design of studies as well as suggesting the optimum analysis of the results. Medical statisticians should be consulted early in the planning of a study. They can contribute in a variety of ways at all stages and not just at the final analysis of the data once the data have been collected.

1.1 Introduction

Most health care practitioners do not carry out medical research. However, if they pride themselves on being up to date then they will definitely be consumers of medical research. It is incumbent on them to be able to discern good studies from bad; to be able to verify whether the conclusions of a study are valid and to understand the limitations of such studies. Evidence-based medicine (EBM) or more comprehensively evidence-based health care (EBHC) requires that health care practitioners consider critically all evidence about whether a treatment works. As Machin and Campbell (2005) point out, this requires the systematic assembly of all available evidence followed by a critical appraisal of this evidence.

A particular example might be a paper describing the results of a clinical trial of a new drug. A physician might read this report to try to decide whether to use the drug on his or her own patients. Since physicians are responsible for the care of their patients, it is their own responsibility to ensure the validity of the report, and its possible generalisation to particular patients. Usually, in the reputable medical press, the reader is to some extent protected from grossly misleading papers by a review process involving both specialist clinical and statistical referees. However, often there is no such protection in the general press or in much of the promotional literature sponsored by self-interested parties. Even in the medical literature, misleading results can get through the refereeing net and no journal offers a guarantee as to the validity of its papers.

The use of statistical methods pervades the medical literature. In a survey of original articles published in three UK journals of general practice; *British Medical Journal* (*General Practice Section*), *British Journal of General Practice* and *Family Practice*; over a 1-year period, Rigby et al (2004) found that 66% used some form of statistical analysis. It appears, therefore, that the majority of papers published in these journals require some statistical knowledge for a complete understanding.

Statistics is not only a discipline in its own right but it is also a fundamental tool for investigation in all biological and medical science. As such, any serious investigator in these fields must have a grasp of the basic principles. With modern computer facilities there is little need for familiarity with the technical details of statistical calculations. However, a health care professional should understand when such calculations are valid, when they are not and how they should be interpreted.

1.2 Why use statistics?

To students schooled in the 'hard' sciences of physics and chemistry it may be difficult to appreciate the variability of biological data. If one repeatedly puts blue litmus paper into acid solutions it turns red 100% of the time, not most (say 95%) of the time. In contrast, if one gives aspirin to a group of people with headaches, not all of them will experience relief. Penicillin was perhaps one of the few 'miracle' cures where the results were so dramatic that little evaluation was required. Absolute certainty in medicine is rare.

Measurements on human subjects rarely give exactly the same results from one occasion to the next. For example, O' Sullivan et al (1999), found that systolic blood pressure in normal healthy children has a wide range, with 95% of children having systolic blood pressures below 130 mmHg when they were resting, rising to 160 mmHg during the school day, and falling to below 130 mmHg at night.

This variability is also inherent in responses to biological hazards. Most people now accept that cigarette smoking causes lung cancer and heart disease, and yet nearly everyone can point to an apparently healthy 80-year-old who has smoked for 60 years without apparent ill effect.

Although it is now known from the report of Doll et al (2004) that about half of all persistent cigarette smokers are killed by their habit, it is usually forgotten that until the 1950s, the cause of the rise in lung cancer deaths was a mystery and commonly associated with diesel fumes. It was not until the carefully designed and statistically analysed case–control and cohort studies of Richard Doll and Austin Bradford Hill and others, that smoking was identified as the true cause. Enstrom and Kabat (2003) have now moved the debate on to whether or not passive smoking causes lung cancer. This is a more difficult question to answer since the association is weaker.

With such variability, it follows that in any comparison made in a medical context, differences are almost bound to occur. These differences may be due to real effects, random variation or both. It is the job of the analyst to decide how much variation should be ascribed to chance, so that any remaining variation can be assumed to be due to a real effect. This is the art of statistics.

1.3 Statistics is about common sense and good design

A well-designed study, poorly analysed, can be rescued by a reanalysis but a poorly designed study is beyond the redemption of even sophisticated statistical manipulation. Many experimenters consult the medical statistician only at the end of the study when the data have been collected. They believe that the job of the statistician is simply to analyse the data, and with powerful computers available, even complex studies with many variables can be easily processed. However, analysis is only part of a statistician's job, and calculation of the final '*p*-value' a minor one at that!

A far more important task for the medical statistician is to ensure that results are comparable and generalisable.

Example from the literature: Fluoridated water supplies

A classic example is the debate as to whether fluorine in the water supply is related to cancer mortality. Burke and Yiamouyannis (1975) considered 10 fluoridated and 10 non-fluoridated towns in the USA. In the fluoridated towns, the cancer mortality rate had increased by 20% between 1950 and 1970, whereas in the non-fluoridated towns the increase was only 10%. From this they concluded that fluoridisation caused cancer. However, Oldham and Newell (1977), in a careful analysis of the changes in age–gender–ethnic structure of the 20 cities between 1950 and 1970, showed that in fact the excess cancer rate in the fluoridated cities increased by only 1% over the 20 years, while in the unfluoridated cities the increase was 4%. They concluded from this that there was no evidence that fluoridisation caused cancer. No statistical significance testing was deemed necessary by these authors, both medical statisticians, even though the paper appeared in a statistical journal!

In the above example age, gender and ethnicity are examples of confounding variables as illustrated in Figure 1.1. In this example, the types of individuals exposed to fluoridation depend on their age, gender and ethnic mix, and these same factors are also known to influence cancer mortality rates. It was established that over the 20 years of the study, fluoridated towns were more likely to be ones where young, white people moved away and these are the people with lower cancer mortality, and so they left behind a higher risk population.

Any observational study that compares populations distinguished by a particular variable (such as a comparison of smokers and non-smokers) and ascribes the differences found in other variables (such as lung cancer rates) to the first variable is open to the charge that the observed differences are in fact due to some other, confounding, variables. Thus, the difference in lung

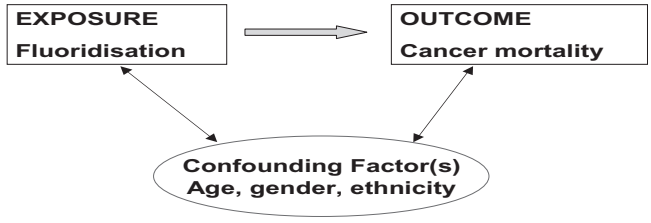

Figure 1.1 Graphical representation of how confounding variables may influence both exposure to fluoridisation and cancer mortality

cancer rates between smokers and non-smokers has been ascribed to genetic factors; that is, some factor that makes people want to smoke also makes them more susceptible to lung cancer. The difficulty with observational studies is that there is an infinite source of confounding variables. An investigator can measure all the variables that seem reasonable to him but a critic can always think of another, unmeasured, variable that just might explain the result. It is only in prospective randomised studies that this logical difficulty is avoided. In randomised studies, where exposure variables (such as alternative treatments) are assigned purely by a chance mechanism, it can be assumed that unmeasured confounding variables are comparable, on average, in the two groups. Unfortunately, in many circumstances it is not possible to randomise the exposure variable as part of the experimental design, as in the case of smoking and lung cancer, and so alternative interpretations are always possible. Observational studies are further discussed in Chapter 12.

1.4 Types of data

Just as a farmer gathers and processes a crop, a statistician gathers and processes data. For this reason the logo for the UK Royal Statistical Society is a sheaf of wheat. Like any farmer who knows instinctively the difference between oats, barley and wheat, a statistician becomes an expert at discerning different types of data. Some sections of this book refer to different data types and so we start by considering these distinctions. Figure 1.2 shows a basic summary of data types, although some data do not fit neatly into these categories.

Example from the literature: Risk factors for endometrial cancer

Table 1.1 gives a typical table reporting baseline characteristics of a set of patients entered into a case–control study which investigated risk factors for endometrial cancer (Xu et al, 2004). We will discuss the different types of data given in this paper.

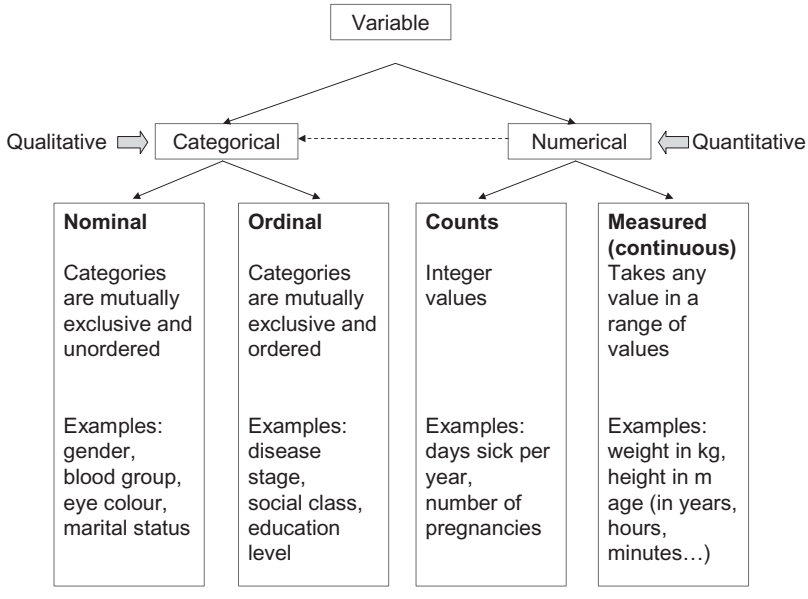

Figure 1.2 Broad classification of the different types of data with examples

Categorical or qualitative data

Nominal categorical data Nominal or categorical data are data that one can *name* and put into categories. They are not measured but simply counted. They often consist of unordered 'either–or' type observations which have two categories and are often know as *binary*. For example: Dead or Alive; Male or Female; Cured or Not Cured; Pregnant or Not Pregnant. In Table 1.1 having a first-degree relative with cancer, or taking regular exercise are binary variables. However, categorical data often can have more that two categories, for example: blood group O, A, B, AB, country of origin, ethnic group or eye colour. In Table 1.1 marital status is of this type. The methods of presentation of nominal data are limited in scope. Thus, Table 1.1 merely gives the number and percentage of people by marital status.

Ordinal data If there are more than two categories of classification it may be possible to order them in some way. For example, after treatment a patient may be either improved, the same or worse; a woman may never have conceived, conceived but spontaneously aborted, or given birth to a live infant. In Table 1.1 education is given in three categories: none or elementary school, middle school, college and above. Thus someone who has been to middle school has more education than someone from elementary school but

Table 1.1 Demographic characteristics and selected risk factors for endometrial cancer. Values are numbers (percentages) unless stated otherwise

Characteristic	Cases	Controls
Number of women (*n*)	832	846
Mean (SD) age (years)	55.3 (8.60)	55.7 (8.58)
Education		
No formal education or just elementary school	204 (24.5)	234 (27.7)
Middle school	503 (60.5)	513 (60.6)
College and above	125 (15.0)	99 (11.7)
Marital status		
Unmarried	14 (1.7)	10 (1.2)
Married or cohabiting	724 (87.0)	742 (87.7)
Separated, divorced, or widowed	94 (11.3)	94 (11.1)
Per capita income in previous year (yuan)		
≤4166.7	230 (27.7)	244 (28.9)
4166.8–6250.3	243 (29.2)	242 (28.6)
6250.4–8333.3	57 (6.9)	50 (5.9)
≥8333.3	301 (36.2)	309 (36.6)
No of pregnancies		
None	62 (7.5)	35 (4.1)
1	137 (16.5)	109 (12.9)
2	199 (23.9)	208 (24.6)
3	194 (23.3)	207 (24.5)
4	141 (17.0)	157 (18.6)
≥5	99 (11.9)	130 (15.4)
Cancer among first degree relatives	289 (34.7)	228 (27.0)
Oral contraceptive use	147 (17.7)	207 (24.5)
Regular exercise	253 (30.4)	287 (33.9)
Age at menarche*	14 (13 to 16)	15 (13 to 16)
Age at menopause (among postmenopausal women)*	50.1 (48.6 to 52.5)	49.4 (47.1 to 51.1)
Body mass index*	25.1 (22.7 to 27.9)	23.7 (21.4 to 26.3)

From Xu et al (2004). Soya food intake and risk of endometrial cancer among Chinese women in Shanghai: population-based case–control study. *British Medical Journal*, **328**, 1285–1291: reproduced by permission of the BMJ Publishing Group.
*Median (25th to 75th centile).

less than someone from college. However, without further knowledge it would be wrong to ascribe a numerical quantity to position; one cannot say that someone who had middle school education is twice as educated as someone who had only elementary school education. This type of data is also known as *ordered categorical data*.

Ranks In some studies it may be appropriate to assign ranks. For example, patients with rheumatoid arthritis may be asked to order their preference for

four dressing aids. Here, although numerical values from 1 to 4 may be assigned to each aid, one cannot treat them as numerical values. They are in fact only codes for best, second best, third choice and worst.

Numerical or quantitative data

Count data Table 1.1 gives details of the number of pregnancies each woman had had, and this is termed count data. Other examples are often counts per unit of time such as the number of deaths in a hospital per year, or the number of attacks of asthma a person has per month. In dentistry, a common measure is the number of decayed, filled or missing teeth (DFM).

Measured or numerical continuous Such data are measurements that can, in theory at least, take any value within a given range. These data contain the most information, and are the ones most commonly used in statistics. Examples of continuous data in Table 1.1 are: age, years of menstruation and body mass index.

However, for simplicity, it is often the case in medicine that continuous data are *dichotomised* to make nominal data. Thus diastolic blood pressure, which is continuous, is converted into hypertension (>90 mmHg) and normotension (≤90 mmHg). This clearly leads to a loss of information. There are two main reasons for doing this. It is easier to describe a population by the proportion of people affected (for example, the proportion of people in the population with hypertension is 10%). Further, one often has to make a decision: if a person has hypertension, then they will get treatment, and this too is easier if the population is grouped.

One can also divide a continuous variable into more than two groups. In Table 1.1 per capita income is a continuous variable and it has been divided into four groups to summarise it, although a better choice may have been to split at the more convenient and memorable intervals of 4000, 6000 and 8000 yuan. The authors give no indication as to why they chose these cut-off points, and a reader has to be very wary to guard against the fact that the cuts may be chosen to make a particular point.

Interval and ratio scales

One can distinguish between *interval* and *ratio* scales. In an *interval* scale, such as body temperature or calendar dates, a difference between two measurements has meaning, but their ratio does not. Consider measuring temperature (in degrees centigrade) then we cannot say that a temperature of 20°C is twice as hot as a temperature of 10°C. In a *ratio* scale, however, such as body weight, a 10% increase implies the same weight increase whether expressed in kilograms or pounds. The crucial difference is that in a ratio

scale, the value of zero has real meaning, whereas in an interval scale, the position of zero is arbitrary.

One difficulty with giving ranks to ordered categorical data is that one cannot assume that the scale is interval. Thus, as we have indicated when discussing ordinal data, one cannot assume that risk of cancer for an individual educated to middle school level, relative to one educated only to primary school level is the same as the risk for someone educated to college level, relative to someone educated to middle school level. Were Xu et al (2004) simply to score the three levels of education as 1, 2 and 3 in their subsequent analysis, then this would imply in some way the intervals have equal weight.

1.5 How a statistician can help

Statistical ideas relevant to good design and analysis are not easy and we would always advise an investigator to seek the advice of a statistician at an early stage of an investigation. Here are some ways the medical statistician might help.

Sample size and power considerations

One of the commonest questions asked of a consulting statistician is: How large should my study be? If the investigator has a reasonable amount of knowledge as to the likely outcome of a study, and potentially large resources of finance and time, then the statistician has tools available to enable a scientific answer to be made to the question. However, the usual scenario is that the investigator has either a grant of a limited size, or limited time, or a limited pool of patients. Nevertheless, given certain assumptions the medical statistician is still able to help. For a given number of patients the probability of obtaining effects of a certain size can be calculated. If the outcome variable is simply success or failure, the statistician will need to know the anticipated percentage of successes in each group so that the difference between them can be judged of potential clinical relevance. If the outcome variable is a quantitative measurement, he will need to know the size of the difference between the two groups, and the expected variability of the measurement. For example, in a survey to see if patients with diabetes have raised blood pressure the medical statistician might say, 'with 100 diabetics and 100 healthy subjects in this survey and a possible difference in blood pressure of 5 mmHg, with standard deviation of 10 mmHg, you have a 20% chance of obtaining a statistically significant result at the 5% level'. (The term 'statistically significant' will be explained in Chapter 7.) This statement means that one would anticipate that in only one study in five of the proposed size would a statistically significant result be obtained. The investigator would then have to

decide whether it was sensible or ethical to conduct a trial with such a small probability of success. One option would be to increase the size of the survey until success (defined as a statistically significant result if a difference of 5 mmHg or more does truly exist) becomes more probable.

Questionnaires

Rigby et al (2004), in their survey of original articles in three UK general practice journals, found that the most common design was that of a cross-sectional or questionnaire survey, with approximately one third of the articles classified as such.

For all but the smallest data sets it is desirable to use a computer for statistical analysis. The responses to a questionnaire will need to be easily coded for computer analysis and a medical statistician may be able to help with this. It is important to ask for help at an early stage so that the questionnaire can be piloted and modified before use in a study. Further details on questionnaire design and surveys are given in Chapter 12.

Choice of sample and of control subjects

The question of whether one has a representative sample is a typical problem faced by statisticians. For example, it used to be believed that migraine was associated with intelligence, perhaps on the grounds that people who used their brains were more likely to get headaches but a subsequent population study failed to reveal any social class gradient and, by implication, any association with intelligence. The fallacy arose because intelligent people were more likely to consult their physician about migraine than the less intelligent.

In many studies an investigator will wish to compare patients suffering from a certain disease with healthy (control) subjects. The choice of the appropriate control population is crucial to a correct interpretation of the results. This is discussed further in Chapter 12.

Design of study

It has been emphasised that design deserves as much consideration as analysis, and a statistician can provide advice on design. In a clinical trial, for example, what is known as a double-blind randomised design is nearly always preferable (see Chapter 13), but not always achievable. If the treatment is an intervention, such as a surgical procedure it might be impossible to prevent individuals knowing which treatment they are receiving but it should be possible to shield their assessors from knowing. We also discuss methods of randomisation and other design issues in Chapter 13.

Laboratory experiments

Medical investigators often appreciate the effect that biological variation has in patients, but overlook or underestimate its presence in the laboratory. In dose–response studies, for example, it is important to assign treatment at random, whether the experimental units are humans, animals or test tubes. A statistician can also advise on quality control of routine laboratory measurements and the measurement of within- and between-observer variation.

Displaying data

A well-chosen figure or graph can summarise the results of a study very concisely. A statistician can help by advising on the best methods of displaying data. For example, when plotting histograms, choice of the group interval can affect the shape of the plotted distribution; with too wide an interval important features of the data will be obscured; too narrow an interval and random variation in the data may distract attention from the shape of the underlying distribution. Advice on displaying data is given in Chapters 2 and 3.

Choice of summary statistics and statistical analysis

The summary statistics used and the analysis undertaken must reflect the basic design of the study and the nature of the data. In some situations, for example, a median is a better measure of location than a mean. (These terms are defined in Chapter 3.) In a matched study, it is important to produce an estimate of the difference between matched pairs, and an estimate of the reliability of that difference. For example, in a study to examine blood pressure measured in a seated patient compared with that measured when he is lying down, it is insufficient simply to report statistics for seated and lying positions separately. The important statistic is the change in blood pressure as the patient changes position and it is the mean and variability of this difference that we are interested in. This is further discussed in Chapter 8. A statistician can advise on the choice of summary statistics, the type of analysis and the presentation of the results.

1.6 Further reading

Swinscow and Campbell (2002) is an introductory text, which concentrates mainly on the analysis of studies, while Bland (2000) and Campbell (2006) are intermediate texts. Altman (1991) and Armitage et al (2002) give lengthier and more detailed accounts. Machin and Campbell (2005) focus on the design, rather than analysis, of medical studies in general.

1.7 Exercises

1. Consider a survey of nurses' opinions of their working conditions. What type of variables are: (i) length of service (ii) staff grade (iii) age (iv) salary (v) number of patients seen in a day (vi) possession of a degree.
2. What differences do you think are there in a discrete measurement such as shoe size, and a discrete measurement such as family size?
3. Many continuous variables are dichotomised to make them easier to understand e.g. obesity (body mass index $>30\,kg/m^2$) and anaemia (haemoglobin level $<10\,g/dl$). What information is lost in this process? If you were told that a patient was anaemic, what further information would you want before treating the patient? How does a label, such as anaemia, help?