1 Uses and Abuses of Medical Statistics

Medical Statistics: A Textbook for the Health Sciences, Fifth Edition. Stephen J. Walters, Michael J. Campbell, and David Machin.

Companion website: www.wiley.com/go/walters/medicalstatistics

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 statisticians should be consulted early in the planning of a study as they can contribute in a variety of ways and not just once all the data have been collected. Thus, medical statistics can influence good research by improving the design of studies as well as suggesting the optimum analysis of the results and their reporting.

1.1 Introduction

Although some health care practitioners may not carry out medical research, they will definitely be consumers of medical research. Thus, 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. The current emphasis on evidence-based medicine (EBM), or more comprehensively evidence-based health care (EBHC), requires that health care practitioners consider critically all evidence about whether a specific treatment works and this requires basic statistical knowledge.

Statistics is not only a discipline in its own right but it is also a fundamental tool for investigation in all biological and medical sciences. 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.

The use of statistical methods pervades the medical literature. In a survey of 305 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 one-year period, Rigby et al. (2004) found that 66% used some form of statistical analysis. Another review by Strasak et al. (2007) of 91 original research articles published in *The New England Journal of Medicine (NEJM)* in 2004 (one of the prestigious peer-reviewed medical journals) found an even higher figure with 95% containing inferential statistics, for example, testing hypotheses and deriving estimates. It appears, therefore, that the majority of papers published in these journals require some statistical knowledge for a complete understanding.

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 seldom give exactly the same results from one occasion to the next. For example, O'Sullivan et al. (1999), found that the systolic blood pressure (SBP) in normal healthy children has a wide range, with 95% of children having SBPs below 130 mmHg when they

were resting, rising to 160 mmHg during the school day, and falling again to below 130 mmHg at night. Furthermore, Hansen et al. (2010) in a study of over 8000 subjects found that increasing variability in blood pressure over 24 hours was a significant and independent predictor of mortality and of cardiovascular and stroke events.

Diagnostic tests are not perfect. Simply because a test for a disease is positive does not mean that the patient necessarily has the disease. Similarly, a negative test does not mean the patient is necessarily disease free. The UK National Health Service invites all women aged 50–70 for breast screening every three years. According to the NHS Breast Screening Information Leaflet (2018, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/840343/Breast_screening_helping_you_decide.pdf): if 100 women have breast screening; 96 will have a normal result and 4 will need more tests. Of these, 1 cancer will be confirmed whilst 3 women will have no cancer detected.

One would think that pathologists, at least, would be consistent. However, a review by Elmore et al. (2017) showed that when it came to diagnosing melanotic skin lesions, in only 83% of cases where a lone pathologist made a diagnosis would the same diagnosis be confirmed by an independent panel. In 8% of cases the lone pathologist would give a worse prognosis, and in 9% of cases they would have underestimated the severity of the disease.

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 many 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 general atmospheric pollution from, for example, exhaust fumes of cars. 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 et al. (2003) moved the debate on to ask whether or not passive smoking causes lung cancer. This is a more difficult question to answer since the association is weaker. However, studies by Cao et al. (2015) have now shown that it is a major health problem and scientists at the International Agency for Research on Cancer (IARC) have concluded that there is sufficient evidence that second-hand smoke causes lung cancer (IARC 2012). Restrictions on smoking in public places have been one consequence and in England and Wales since 1 October 2015 it has been illegal to smoke in a vehicle carrying anyone under the age of 18.

With such variability, it follows that in any comparison made in a medical context, such as people on different treatments, differences are almost bound to occur. These differences may be due to real effects, random variation or variation in some other factor that may affect an outcome. It is the job of the analyst to decide how much variation should be ascribed to chance or other factors, 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 – Drinking Coffee and Cancer (IARC 2018)

In 2016, a working group of 23 scientists from 10 countries met at IARC in Lyon, France, to review the research evidence of whether or not drinking coffee is carcinogenic and causes cancer. They reviewed the available data from more than 1000 observational and experimental studies. In rating the evidence, the working group gave the greatest weight to well-conducted studies that controlled satisfactorily for important potential confounders, including tobacco and alcohol consumption. For bladder cancer, they found no consistent evidence of an association with drinking coffee, or of a dose–response relationship, that is drinking more coffee increased the incidence of cancer. In several studies, the relative risks of cancer for those drinking coffee compared to non-drinkers were increased in men but women were either not affected or the risk decreased. IARC (2018) concluded from this that there was no evidence that drinking coffee caused bladder cancer and, as Loomis et al. (2016) stated 'that positive associations reported in some studies could have been due to inadequate control for tobacco smoking, which can be strongly associated with heavy coffee drinking'.

In the above example tobacco and alcohol consumption are examples of confounding variables as illustrated in Figure 1.1. In this example, the individuals exposed or drinking coffee are typified by their tobacco and alcohol consumption, and these same factors are also known to influence cancer incidence rates.

Any observational study that compares populations distinguished by a particular variable (such as a comparison of coffee drinkers and non-coffee drinkers) and ascribes the differences found in other variables (such as bladder 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 bladder cancer rates between coffee drinkers and non-drinkers has been ascribed to genetic factors; that is, some factor that makes people want to drink coffee also makes them more susceptible bladder cancer. The difficulty with observational studies is that there is an infinite source of potential 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 trials, where the alternative interventions (the exposure variables) 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

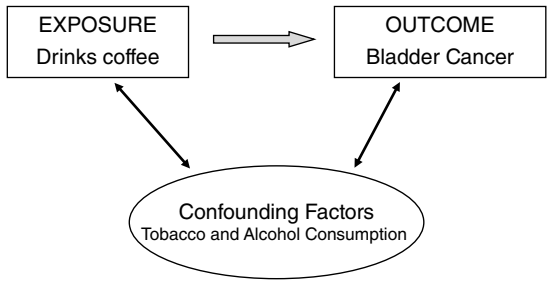

Figure 1.1 Graphical representation of how confounding variables may influence both exposure (drinking coffee) and bladder cancer incidence.

the exposure variable as part of the experimental design, as in the case of drinking coffee and bladder cancer, and so alternative interpretations are always possible. Observational studies are further discussed in Chapter 14.

1.4 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 research 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, the statistician 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 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 6.) This statement means that one would anticipate that in only one study in five (20%) 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 survey 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 14.

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, perhaps, because intelligent people were more likely than the less intelligent to consult their physician about migraine.

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 14.

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 15), 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 15.

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 Chapter 2.

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 2.) 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 or she 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 7. A statistician can advise on the choice of summary statistics, the type of analysis and the presentation of the results.

Medical Statistics and Data Science

Because of the availability of large amounts of data over the last few decades, the term data science has emerged to describe the substantial current intellectual effort around research with the goal of extracting information from these data. The type of data currently available in all sorts of application domains is often massive in size, very heterogeneous and far from being collected under designed or controlled experimental conditions. Nonetheless, it contains information, often substantial information, and it has been argued that data science is a new interdisciplinary approach that makes maximal use of this information. However, data alone is typically not that informative and (machine) learning from data needs conceptual frameworks. Data science would seem to encompass statistics. However, we would argue that statistics is crucial for providing conceptual frameworks that enhance the understanding of fundamental phenomena, highlight limitations and provide a formalism for properly founded data analysis, information extraction and quantification of uncertainty, as well as for the analysis and development of algorithms that carry out these key tasks.

As taught at a number of universities, data science differs from statistics in a number of ways. Statistics originated before the computer and its core concern is with statistical models. However, no serious statistician is beguiled into confusing their model with reality ('All models are wrong, but some are useful' to quote the famous statistician John Tukey). However, models are very useful in describing how the world might be, and for making generalisations beyond the data. Data science is empirical, reliant on large data sets, whereas one of the key successes of statistics is doing inference on relatively small data sets, such as those available in agriculture and laboratories. Data science is often used for prediction, and the idea is that with the vast amounts of data now available electronically (such as that provided by national health services) one can look at empirical relationships and build up accurate predictors, such as how drugs will behave in individuals. These predictions are often highly successful, but lacking models it can be difficult to know why it makes some predictions, and how generalizable the predictions might be. Data science is related to the concept of 'big data'. However, simply because a sample is large does not mean it is unbiased.

A case in point is the reported link between taking hormone replacement therapy (HRT) and *lower* heart disease rates observed in some large data sets. However, a key issue is whether women who use HRT are already more health conscious. It can be difficult to know whether this fact is adequately accounted for in conclusions drawn from the big data. Thus, it was only when the results of the randomised controlled trial of the use of HRT (Writing Group for the Women's Health Initiative Investigators 2002) became available that HRT was shown not to protect against heart disease. In fact, the trial identified an increased risk for total cardiovascular disease with hazard ratio 1.22 and 95% confidence interval 1.09 to 1.36 (the technical terms will be explained in Chapter 11). In this example, big data led to a wrong conclusion.