Fear or favour? Statistics in pathology

Angie Wade

Statistics have a role to play in most areas of medical research including the field of pathology. We have come a long way since 1954 when the *British Medical Journal* published excerpts from a debate held by the Study Circle on Medical Statistics as to whether the then growing influence of statistics in medicine was, in fact, welcome.¹ One speaker declared that, "medicine was an art, statistics a science; he conceded that the latter had its uses, but when it came to mixing science and art, statistics was as out of place as a skillet in a Crown Derby tea-service." He concluded that "statistics might be all very well for the elite but were a menace to the mob." Someone else "referred darkly to the deliberate misuse of statistics, fostered—for what purpose?—by statisticians themselves. Statistical publications, he said, could be recognised by the prolixity of their tables. In his view no papers should contain any tables at all." The debate concluded with the motion that the influence of statistics should be welcomed in all branches of medicine and this was carried by a narrow majority on a show of hands.

In the intervening 45 years there has been a mushrooming of statistical literature designed to assist the medical researcher, with numerous articles highlighting misuses of statistics and giving pointers towards improvement. There has been a growing understanding that statisticians are concerned with the whole process of research, from study design through to final conclusions, and are not merely purveyors of p values and analytical methodology. The more recent evidence based medicine movement has served to further publicise this recognition. As H G Wells predicted, "statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

The lessons have been numerous and only the major developments are reviewed here. Emphasis has increasingly been placed on identifying a well defined and answerable research question before undertaking any study. This seemingly obvious prerequisite may be the hardest part, finding the right question often being more troublesome than finding the right answer. In the words of Einstein, "The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill." The incorporation of suitable control groups and some quantification, before starting a study, of the necessary sample size required to conclusively answer the research question have been stressed. The need for some formal statistical comparison of the results, usually resulting in a p value, has also been encouraged. An informal review of the publications in this journal over the last 20 years shows that this latter

point has indeed been taken on board. The majority of the papers published in *JCP* in 1978 contained little or no formal statistical analysis: during the whole of that year only 38 papers contained any p values and most of those had only one. By contrast, in the current editions most submissions have at least some formal analysis and usually contain several p values. When interpreting the clinical value of results statisticians have, in more recent years, stressed the importance of quantifying the size of any effect, rather than merely relying on a significance level as given by a p value. To this end, the current guidelines for this journal state that "95% confidence intervals should be used wherever appropriate." This has certainly led to a dramatic increase in usage. During the whole of 1978 there was only one confidence interval presented in *JCP*, compared with the current situation where they are to be found in most issues.

So, in common with most others, this journal has seen a secular trend in the use of statistics, and the statistical quality of published research is undoubtedly superior to that seen 20 years ago. As we enter the new millennium is there anything that can be done to assist yet further improvements?

One potential barrier to such improvement is the dearth of statistical guidance aimed specifically at pathologists. It may seem strange to suggest that the wealth of published statistical literature is not directly applicable to research in the field of pathology. Every discipline tends to use certain types of study design and forms of statistical analyses more than others. The necessary information is out there but it may be somewhat off-putting to have to delve through a mountain of irrelevant material to find it. By identifying the areas of main interest we can considerably reduce the ground that must be covered to gain the necessary knowledge to produce good quality research that answers useful questions in our particular area. For example, the majority of medical statistics texts aimed at the non-statistician urge us to perform randomised controlled trials, anything else being considered inferior, yet these are hardly ever used by and are largely irrelevant to pathologists.

Most of the research questions addressed by studies in this journal concern the comparison between two or more previously defined groups (for example, diseased and healthy, different diseased groups, or those at different stages of the same disease); assessing the reliability/validity/reproducibility of measurements; predicting time to death/recovery/relapse/infection or using new measurements to improve diagnostic accuracy.

Department of
Epidemiology and
Biostatistics, Institute
of Child Health, 30
Guilford St, London
WC1N 1EH, UK
A Wade

Correspondence to:
Dr Wade

The rest of this paper will focus on two aspects of medical statistics that are relevant to all of these scenarios yet are still widely misunderstood or poorly presented. The aim is to provide a further learning brick to build on the improvements that have already been seen over the last two decades.

Choice of sample(s)

INDIVIDUALS

When we perform studies we are trying to find out what happens in the population. For example, do two measuring instruments give the same readings when applied to patients in disease group X? Can we accurately and consistently measure Y? Can we predict survival or time to relapse from P, Q, and Z? We cannot measure the whole population, so we observe a subset or sample of individuals and from these infer what we think happens in the population as a whole. Statistical analyses are used to make this inference.

Studies can be irretrievably ruined by the biased choice of samples, particularly if we are unaware of the size and direction of any bias. However, choice of sample appears to receive little thought and is often made according for convenience rather than representativeness. Commonly all available samples from a given laboratory or hospital may be included. While this is all that may be feasible within the time and practicality constraints imposed on the researchers, there should be some attempt to identify whether this sample is in fact representative of the population of interest. For example, does this hospital tend to get referred patients at all stages of this disease or is it biased towards the more symptomatic? Is the area which this laboratory/hospital serves socially and ethnically representative? Given this kind of information, the reader can decide whether the results are likely to apply to their own population.

Quite commonly it is a subgroup of patients over a certain time frame who are included, these being chosen according to availability of blocks, tissue samples, or data. In this case there needs to be some discussion of the representativeness of the subgroup. For example, were those with available tissue more severely ill? Do they tend to have different underlying diseases? Were data more carefully recorded in the more unusual diagnostic cases?

Control groups which consist of "healthy volunteers" may be used. This group should ideally be similar to the disease group except for the presence of disease. It is of interest to know precisely how this group has been recruited to help determine whether it constitutes a reasonable comparison group. For example, sometimes laboratory staff or patients admitted for reasons unrelated to the present study interest are used as controls. In fact either of these might be considered unsuitable. The former, laboratory staff, may be younger than the study disease group and the latter may not be entirely normal with respect to the study measures.

Volunteers, whether from the disease or the control group, may differ from non-volunteers,

and it will usually be impossible to assess the extent of that difference. It is best to avoid advertising for volunteers as a means of recruitment. If 50% of those approached to participate refuse then at least we have some measure of how representative the final sample is.

When the sample is to consist of a subset of some larger group of eligible individuals, then this subset should be randomly chosen, that is in a manner unbiased by the characteristics of the individuals and in a non-systematic way. Random selections *must* be made using either tables or suitable software and this should be made explicit in the description of the sample selection process.

SPECIMENS WITHIN INDIVIDUALS

Having identified individuals to be included in the study, there may be the further selection of a particular specimen to be analysed. This has been shown² to be "the most observer dependent and therefore most subjective step." Many studies state that "representative sections" or "systematically selected areas" were chosen, but precisely how this was done is not made explicit. If there is a system it should be clearly outlined so that others can make comparable choices. If selection is random, the methodology should be specified.

To summarise, while the ideal of comparable and representative samples from the study groups concerned (for example, disease X and healthy controls, disease X and disease Y) is rarely attained, we can improve published studies by giving full details of the selection process for both the individuals and specimens from those individuals. The representativeness of these should be discussed to enable readers to identify applicability and potentially confounding variables.

A further and related point is that comparability can be improved by ensuring that those who make the assessments are blind to the study group. Where several assessors are involved then there should be high inter-rater reliability.

Size of sample(s)

Samples are used to estimate population effects. It is intuitively obvious that a larger sample will give a more precise estimation of the population value. For example, if 20% of the population display trait X then in a sample of 10 from this population we would not be surprised to find anywhere between 0 and 5 individuals with the trait (0–50%). If we sample 100 individuals we would not be surprised to observe anywhere between 13 and 29 (13–29%) with the trait; more extreme numbers may lead us to doubt whether the true prevalence of X is actually 20%. The thinking is similar when we present confidence intervals with sample estimates. From our sample we estimate the population value (for example, the mean or proportion; the difference in means or proportions between normal and diseased individuals or the median survival in different groups). We do not expect this estimate to be exact, although we know that the larger the

sample the more precise it will be. A confidence interval gives the range of population values (or differences) that our sample(s) are compatible with: 95% confidence intervals give the range within which we are 95% confident the population value lies. Clearly the addition of a confidence interval facilitates the clinical interpretation of the results and highlights any limitations caused by sample size. Most statistical packages now give confidence intervals as standard; details of calculation can be found elsewhere.^{3,4}

The power of a study is its ability to detect a difference of a given size.⁵ For example, suppose QRZ in the population of individuals with disease K is on average 10 and always varies between 5 and 20 compared with an average of 20 and range of 10–35 for normal individuals. (Note that the ranges are not symmetric around the averages and this is to stress the idea that this thinking applies not only to normally or symmetrically distributed data.) We may by chance randomly sample disease K patients with a tendency to higher values and normal controls with a tendency to lower values and hence there will be no significant difference in the average values in the samples. We may therefore wrongly conclude that there is no difference in average QRZ between the groups. Of course this will not always happen and how often it does depends on both the sample sizes (the larger the samples the more closely they tend to approximate their respective population means) and the variability of the measurements in the populations (if disease K measures are mostly between 8 and 12 and the normals between 18 and 22, then it will be less likely to happen than if the values of QRZ are more evenly spread across the range in each group). The power of the study, usually expressed as a percentage, tells us how often a given difference will be detected for a certain variability and sample size. As the variability of a measure is fixed (that is, it exists in the population and there is nothing we can do about it except perhaps choose a more homogeneous population which will change the research question), the aim is to choose a sample size that will detect a clinically important difference with reasonable power. Power is usually set at 80% or above. A value of 80% means that four times out of five the study will detect the difference if it exists.

It is now accepted as standard practice that all published randomised controlled trials should include some statement regarding the power of the study.⁶ It is less well recognised that a similar proviso would benefit the presentation of all studies, including non-randomised and single group descriptive studies. Such a policy safeguards the researcher against wast-

ing time with a sample that is too small to give conclusive answers. It also serves to assist interpretation where results are non-significant, in which case we want to know the power that the study had to detect a difference. The combination of no power calculations, p value reporting, and interpretation with little or no use of confidence intervals is a recipe for potential disaster.

Kirkwood⁷ gives a good overview of the most commonly used power and precision calculations. Other useful references⁸ are for ordinal variables,^{9,10} reliability coefficients,^{4,11–13} survival analyses,^{14,15} and for testing equivalence^{16,17} (which requires larger samples than to show a difference).

Conclusion

If we are to have a new year's resolution for statistics in pathology let it be that we will use unbiased sample selection methods which are fully reported, perform power calculations, and present results with confidence intervals. In this way we can ensure that we are selecting and interpreting our data without fear or favour of being misled by biased samples or mystical p values. Conclusions based solely on the latter should be valued forthwith.

- 1 Numbering Off. *BMJ* 1954;iii:1314.
- 2 Belien JAM, Somi S, de Jong JS, *et al*. Fully automated microvessel counting and hot spot selection by image processing of whole tumour sections in invasive breast cancer. *J Clin Pathol* 1999;52:184–92.
- 3 Gardner MJ, Altman DG. *Statistics with confidence—confidence intervals and statistical guidelines*. London: BMJ Publishing Group, 1989.
- 4 Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the Kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Stat Med* 1992;11:1511–19.
- 5 Florey C du V. Sample size for beginners. *BMJ* 1993;306:1181–4.
- 6 Altman DG. Better reporting of randomised controlled trials: the CONSORT statement. *BMJ* 1996;313:570–1.
- 7 Kirkwood BR. Calculation of required sample size. In: *Essentials of medical statistics*. Oxford: Blackwell Scientific Publications, 1988: chapter 26.
- 8 Daly LE. Confidence intervals and sample sizes: don't throw out all your old sample size tables. *BMJ* 1991;302:333–6.
- 9 Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995;311:1145–8.
- 10 Julious SA, Campbell MJ. Sample size calculations for paired or matched ordinal data. *Stat Med* 1998;17:1635–42.
- 11 Donner A, Eliasziw E. Sample size requirements for reliability studies. *Stat Med* 1987;6:441–8.
- 12 Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. *Stat Med* 1998;17:101–10.
- 13 Donner A. Sample size requirements for the comparison of two or more coefficients of inter-observer agreement. *Stat Med* 1998;17:1157–68.
- 14 Machin D, Campbell MJ. Comparing two survival curves (logrank comparisons). In: *Statistical tables for the design of clinical trial*. Oxford: Blackwell Scientific Publications, 1987: chapter 9.
- 15 Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1982;1:121–9.
- 16 Jones B, Jarvis P, Lewis JA, *et al*. Trials to assess equivalence: the importance of rigorous methods. *BMJ* 1996;313:36–9.
- 17 Hauschke D, Kieser M, Diletti E, *et al*. Sample size determination for proving equivalence based on the ratio of two means for normally distributed data. *Stat Med* 1999;18:93–105.