Teaching Clinical Statistics in the Medical Curriculum

Geoffrey Berry, P Glasziou and J M Simpson - Sydney, Australia

1. Introduction

It is widely recognised that medical students should be taught medical statistics, since it is part of the curriculum in many undergraduate medicine programmes. In spite of this a constant challenge in teaching medical statistics is to convince the medical students that the subject is relevant to their future careers. If it is accepted that the justification for including medical statistics in the undergraduate medicine curriculum is its relevance in medical practice, then it follows that the course content should concentrate on those parts that are relevant to clinical practice, and it might be useful to emphasise this by naming that part of medical statistics as clinical statistics.

What is relevant depends on what medical practitioners do. Some follow careers as research workers, or engage themselves in clinical research as an extension of clinical practice. The proportion of doctors carrying out research is unknown. In 1988 there were 12493 doctors registered in New South Wales; of these, 5068 (41%) were in general practice, 3768 (30%) were specialists, and 1150 (9%) were registrars (Medical Registration Board of New South Wales, personal communication). Only 193 (1.5%) listed their main activity as research or teaching, but a proportion of the specialists and registrars would be involved in some research. Since registration is required mainly for clinical practice some doctors involved in full-time non-clinical research might not be registered, but even so the proportion of trained doctors engaging in research is small. On the other hand at least 80% of those registered were practising clinicians who would be required to make decisions regularly on diagnosis and treatment.

Undergraduate medical training is just the first step in the acquisition of medical knowledge and skills. During a medical career retraining is a constant process as the practitioner keeps up-to-date with developments in his or her area of medicine. The majority of such developments are described in research papers, and in many of these statistical methods will have been used. It is thus important that all medical practitioners have some knowledge of the statistical concepts used in medical research. In particular they need to be able to interpret results expressed in statistical terms, and this means having a good understanding of what is meant by statistical significance and what a confidence interval represents. They need knowledge of the strengths and weaknesses of the different study types, cohort study, case-control study and randomised controlled trial, and of the superiority of the last of these types where it is feasible. The details of the methods such as the t or chi-squared tests are less important and necessary mainly as illustration of the application of the more general concepts.

Of course the minority who go on to medical research need much more but acquisition of this extra knowledge and skill would be more appropriately carried out at the beginning of their research work when its application is near at hand and motivation is therefore higher.

Since most medical practitioners are making medical decisions regularly it is necessary to consider what statistical methods are helpful for such decisions. It has long been recognised that Bayes' theorem has a place here; undoubtedly a method which adjusts prior probabilities (of different diagnoses) to take account of new data (physical examinations and diagnostic tests) seems tailor-made for the diagnostic process. Bayes' theorem is just one tool used in decision making and it would be appropriate that medical practitioners had a wider knowledge of the methods and application of decision making.

An important reason for emphasising the use of Bayes' theorem is that taking proper account of the prevalence of disease in determining the predictive value of a diagnostic test is far from intuitive. There is evidence that the positive predictive value will be grossly overestimated for a test with high sensitivity and specificity applied in a situation where the disease is uncommon. For example, for a test for cancer with sensitivity of 79.2% and specificity of 90.4% applied to a patient with prior probability of cancer of 1%, the majority of physicians in an informal sample estimated the positive predictive value to be about 75%, whereas the correct value by Bayes' theorem is 7.7% (Eddy, 1982). It appeared that the physicians were assuming that P(cancer|positive test) was the same as P(positive test|cancer). The potential for mismanagement of patients based on an incorrect perception of the diagnostic information provided by diagnostic tests is enormous.

2. Current courses in Australian medical schools

Ten Australian universities have medical faculties or schools with an undergraduate medicine course leading to qualification as a medical practitioner. These ten medical courses have an annual intake of about 1400 students (range 50 to 250). In 1989 each medical school was invited to participate in an exchange of information on the scope of medical statistics in the medicine programme. Information was requested on organisation of the teaching, curriculum, timing, method of teaching, number of students and method of examination or assessment. The academic staff responsible for medical statistics responded from all ten schools and the information was collated (Berry, 1989).

In five schools medical statistics was taught in the first year, in one school in the second year, in two in the third year, whilst in the other two the teaching was spread over the first three years. The number of hours available for teaching varies widely between 6 and 35 hours. In almost all the courses the teaching method is a mixture of lectures and tutorials or practical classes. Five of the courses are completely separate, although in one of these the practical classes are held in collaboration with practicals on physiology of the blood and respiration. The other five courses are a component of a larger course, in three cases in community medicine, one in behavioural science, and one in physiology.

The curricula were described with different amounts of detail so that it is impossible to give definite figures on the number of schools including various items. Nevertheless all ten schools included the concept of significance testing, and in at least eight the t and chi-squared tests were covered. In six cases it was clear that estimation was given some emphasis as well as hypothesis testing. In seven cases regression or correlation was mentioned, in four non-parametric tests, and three mentioned the concept of power and sample size determination. In only five of the curriculum summaries was probability mentioned explicitly, and in two of these it was stated that Bayes' theorem was covered. In one case decision making was mentioned.

All courses appear to cover statistical methods used in medical research to the extent necessary for critical appraisal of the literature, but the application of probability, in particular Bayes' theorem, to medical decision making is only rarely mentioned.

3. Proposed curriculum

The course given in the undergraduate medicine curriculum at Sydney University over the last eight years has included the application of Bayes' theorem but its main emphasis has been on methods used in research. The curriculum is currently being revised and decision making will be given more emphasis from 1991. The objectives which we consider it is important to include are categorised in two broad areas:

(i) to develop skills in the probabilistic calculations relevant to clinical practice;
(ii) to develop an understanding of statistical considerations in medical research.

The first set of skills is to help students become more rational decision makers when faced with uncertainty, and the second to help them become more discerning readers of medical literature. We may break these two areas into the following components:

(i) Students should develop skills in calculating probabilities in clinical situations, and be able to base decisions on these probabilities.

Sources of variability in clinical data: Students need to be aware of the variety of meanings to the "normal" range, e.g. is it defined by disease, by risk, or by a 95% range of the population? They also need to understand the sources of variation in clinical measurements: technical error, biological fluctuation, and observer (dis)agreement.

Diagnostic tests: Sensitivity and specificity provide a sufficient characterisation of a diagnostic test and are approximately invariant over a variety of populations. Students should be able to calculate sensitivity and specificity from research papers. They should also understand the underlying principle of the trade-off between the two. However, receiver operating characteristics (ROC) curves would not be covered.

Bayes' theorem: Ignoring prior information is a common fault in people's reasoning under uncertainty. Both the calculations and the pitfalls of probabilistic reasoning need to be well understood by future clinicians whose everyday practice will involve imperfect diagnostic tests. In particular, students need to be attuned to the importance of prior probability, e.g. how the predictive value changes between diagnostic and screening uses of a test. Furthermore, it needs to be made clear that symptoms and signs are also forms of diagnostic tests and that the same principles apply.

Likelihood ratios: To enable the use of Bayesian calculations in practice, the Bayes' nomogram developed by Fagan (1975) considerably reduces the effort involved. The sensitivity and specificity are converted to two likelihood ratios. Using the nomogram, predictive values can be derived in seconds with the aid of a ruler. The method generalises easily to multi-level tests and is feasible for use on medical wards.

Expected values: Much of what doctors do involves rational gambling, e.g. if a patient has a chance p of appendicitis, we must guess the expected mortality if we (i) operate now, or (ii) wait four hours and operate if there is no improvement. The expected mortality, which will depend on the prior chance p, can be used as a basis for decision making (Weinstein and Fineberg, 1980). This is sometimes difficult for clinicians to admit to, and can lead to inappropriate strategies which attempt to avoid uncertainty: "We can't take any chances" being a common dictum. This may lead to both over-testing and over- treatment.

Probability thresholds: Students should be able to see whether a decision based on expected values is robust, and be able to calculate simple probability thresholds to study this. An example is: if the probability of appendicitis is low, then waiting is best; if it is high, then immediate operation is best; and between these there is a threshold probability where the two strategies have equal expected mortalities. These thresholds can be quite simply calculated with Fagan's nomogram.

(ii) Students should have the ability to interpret (and criticise) statistical aspects of results presented in research reports. They should be able to see when statistical methods are appropriate, and whether an appropriate general method has been applied, e.g. if a hypothesis is suggested, is this tested; if an estimate is called for, is this unbiased and is a confidence interval given?

Random sampling techniques: Students should understand the methods and purpose of simple random sampling, and the difference between bias and noise. They should become "critical" readers of survey figures in their general reading, e.g. in medical texts and newspapers, asking how the sample was chosen, was it large enough, and are the results presented appropriately? Huff (1954) provides many good examples of the types of misuse and poor sampling that we would expect students to be able to detect.

Confidence intervals and P-values: Students should be able to interpret both a confidence interval and a P-value in the context in which they are quoted, and recognise an appropriate use or lack thereof. In teaching this, we should foster an intuition for the uncertainty involved in experiments by getting students to guess the confidence intervals before they calculate them. They should understand that "truth" is not replicated in every sample.

Type I and Type II errors: Students need to be aware that studies do not always give the "correct" answers. In particular, they need to be aware of power considerations in negative studies, and the importance of sample size. This leads naturally on to meta-analysis which could be mentioned but without the numerical methods being taught.

Randomisation: The idea of bias and noise, dealt with in sampling, should be extended to comparative studies, and some elementary causes of bias discussed. Students should be aware that the best way to minimise confounding is by randomisation, and that a consistent and blinded assessment minimises measurement bias. They should also understand what randomisation is not, e.g. not an alternation of treatment and control. The rationale for an "intention-to-treat" analysis needs to be covered.

Regression to the mean: Students should develop an intuition for when this is likely to occur in clinical situations, e.g. repeating a single high blood pressure reading usually leads to a lower reading. This phenomenon often applies to the natural history of illnesses, e.g. a patient with more frequent migraines recently is likely to improve if this is just part of a pattern of fluctuating frequencies. Regression to the mean is often an explanation for superficially favourable results in single arm trials which show patients "improved", and hence indicates that a control group is needed to see if the treated group "improved" to a greater extent. Regression to the mean is also often an explanation for claims that a treatment is most effective in the worst patients.

4. Discussion

It has been argued that decision making is an integral part of a physician's work and hence that the tools for this process need to be included in an undergraduate medical course. Frequently the decision making is probability based and other decisions are consequent on the probability. For example, based on signs, symptoms and an examination, a physician might make a provisional diagnosis of appendicitis. But the probability attached to the diagnosis influences the next decision: operate immediately, wait and see, or obtain a plain abdominal x-ray. The usefulness of the last course depends on the sensitivity (0.4) and the specificity (0.9) of the x-ray (Sox et al., 1988). If the probability of the diagnosis is moderately high then a negative x-ray, with such a low sensitivity, does little to rule out the diagnosis. So ordering the x-ray would be disadvantageous to such patients. This illustrates the importance of calculating the posterior probabilities using Bayes' theorem before ordering a test to see if the result of a test is likely to influence the treatment decision; if not, the test is at best a waste of resources and at worst may lead to delay in implementing the proper treatment.

The possible usefulness of a test may be appreciated using two mnemonics coined by Sackett (personal communication):

> SnNout For a test with high sensitivity (Sn) a negative (N) result rules the diagnosis out (out).
>
> SpPin For a test with high specificity (Sp) a positive (P) result rules the diagnosis in (in).

Appreciation of these principles underlines that a test does not necessarily have to have both a high sensitivity and a high specificity to be useful. A test should not be regarded as universally good or bad but its usefulness depends on the circumstances in which it is used. A test that is very useful in a tertiary referral hospital, where the prior probability of the disease is moderate, may be of little use in a population screening situation, where the probability of disease is low.

The majority of medical statisticians are involved in medical research, and are consulted by medical practitioners carrying out research. They are probably consulted only very occasionally, if ever, on problems to do with decision making in the diagnostic area. This may be part of the reason why the methods of medical research seem to dominate the curricula; medical statisticians tend to teach those areas that seem important based on their experience in interacting with physicians whom they or their predecessors taught ten to twenty years earlier. There are many texts on statistical research methods but few on the application of statistics to medical practice; the book by Ingelfinger et al. (1987) is a notable exception, although at too high a level for an introductory course. Perhaps statisticians have failed to heed the warning they teach; just as patients in a teaching hospital are unrepresentative of the community, so also physicians engaging in research are unrepresentative of the total group of physicians. The problem is not, however, as simple as that. Doctors involved in research usually realise the necessity of using statistical methods and take steps to acquire or "hire" the necessary expertise. But those involved in diagnostic decision making are quite likely to continue with little realisation that there are statistical methods that can help. The challenge is to supply the methodology in the medical curriculum in such a way that it will be applied in future practice, and so lead to more efficient diagnosis, treatment and use of resources.

References

Berry, G (ed) (1989) *The Teaching of Medical Statistics at Undergraduate and Postgraduate Levels in Australian Medical Schools.* Department of Public Health, University of Sydney.

Eddy, D M (1982) Probabilistic reasoning in clinical medicine : problems and opportunities. In: D Kahneman, P Slovic and A Tversky (eds) *Judgement Under Uncertainty : Heuristics and Biases.* Cambridge University Press, Cambridge, 249-267.

Fagan, T J (1975) Nomograms for Bayes' formula. *N.Engl.J.Med.* **293**, 257.

Huff, D (1954) *How to Lie with Statistics.* Penguin Books.

Ingelfinger, J A, Mosteller, F, Thibodeau, L A and Ware, J H (1987) *Biostatistics in Clinical Medicine* (2nd ed). MacMillan, New York.

Sox, H C, Blatt, M A, Higgins, M C and Marton, K I (1988) *Medical Decision Making.* Butterworths, Boston.

Weinstein, M C and Fineberg, H V (1980) *Clinical Decision Analysis.* W B Saunders, Philadelphia.