Commentary

Confidence limits and the limits of confidence

F. KEE

From the Department of Epidemiology and Public Health, The Queen’s University of Belfast, Belfast, UK

Introduction

Someone once said that what goes up, must come down, but this never seems to be the case with workload. While most busy clinicians have been doing their best to keep their heads above water, others have been temporarily distracted by the many column inches devoted to whether or not Evidence-Based Medicine represents a new paradigm for their clinical practice. At times the debate has verged on the bilious, each side attempting to claim the high ground of authority, either by dint of training or experience. The view from one camp, for example, is that ‘... the difficulty with evidence based medicine lies with its exponents—theyir arrogance, their jargon and their penchant for denigrating others.’1 At its more extreme, statisticians should not evaluate clinical data, full stop. Witness the correspondence in the BMJ of September 26th 1998: ‘... We believe that the paper by the Cochrane injuries Group shows ignorance and incomprehension on the part of statisticians untrained in burns care ...’2

An eminently more sensible rejoinder from the pages of the Lancet, however, reminds us that ‘... knowledge based on a scientific discourse is democratic and open to debate and knowledge based on expertise is oligarchic and closed ...’3

A more sanguine Professor Sackett, EBM’s foremost exponent, views clinical expertise as the safety net that will save us from a slide towards cookbook medicine: ‘External clinical evidence can inform, but can never replace individual expertise and it is this expertise that decides whether the external evidence applies to the individual patient at all and, if so, how it should be integrated into a clinical decision.’4

This debate will prove sterile without a little collective humility. Alvan Feinstein has recently reminded the profession of the words of George Santayana that ‘... those who cannot remember the past, are condemned to repeat it.’5 A moment’s reflection on the early controversies in medical statistics in the first decades of the nineteenth century might help us recognize some of the unlearnt lessons and the interdependence between evidence and belief that might also expose the frailty of clinical expertise.

Confidence limits

Pierre Charles Alexander Louis (1787–72) is widely credited as the forefather of clinical epidemiology.6 His work was influential in highlighting the importance of a numerical analysis of bedside practice. For more than 10 years after completing his doctoral dissertation (on tuberculosis), he meticulously collected his observations from the wards of la Charité and the dissecting room before publishing his ‘Recherches sur les effets de la saignée’ in 1835.7 He had observed, for example, that in patients with ‘pneumonitis’ the death rate after blood letting was 18/47 but only 9/36 in those who were not bled. He was alert both to the possibility of confounding by covariates like age and by intention to treat.

Although never claiming statistical credentials, he drew on popular expositions of the ‘calculus of probability’ in texts such as those of Laplace, invoking the authority of mathematical reasoning: ‘It is impossible to appreciate each case with mathematical exactness and it is precisely on this account that

Address correspondence to Professor F. Kee, Department of Epidemiology and Public Health, The Queen’s University of Belfast, Mulhouse Building, Grosvenor Rd, Belfast, BT12 6BJ. e-mail: f.kee@qub.ac.uk

© Association of Physicians 2000
The Marquis de Condorcet, was developing concepts now de rigeur or proofs admissible in a court of law. Poisson, in variety of judgement biases.

Concise Oxford Dictionary describes evidence as available evidence. (my italics) or proposition or indicating poorly understood, but there is ample evidence that supporting evidence statistics, medicine can never be anything but a individual patient) can be a sort of safety net results could vary between 26% and 49%, though as 52/140 (37%), Gavarret demonstrated that the fever death rate observed by Louis. Quoted originally amounted to 99% confidence limits for the typhoid Poisson's formulae led to the calculation of what medical tact, of a kind of divining power of Poisson and derived the likely 'limits of oscillation' foretells a lesson that the profession has been rather working with a colleague of Louis's, Gabriel Andral. bution, urging their colleagues to 'avoid the illusions Polytechnique in 1829 to become an artillery officer, the variability incumbent on sample size. Enthusiastic Statistique Médicale

have strictly paralleled that of a modern frequentist A burgeoning literature is now casting new light on his conception of these limits would probably not be disregarded without sensibly affecting the exactness of the results.' However, he was soon in conflict with many of the leading authorities of the day. His arch critics were Victor Broussais, the Chief Physician at the Parisian military hospital the Val de Grace, and Francois Double. The latter maintained that to apply numerical analysis to clinical practice it would be necessary to 'strip away the individual in order to arrive at the elimination of all which the individual would have been able to introduce accidental to the question' and that, (with a certain resonance for today's debate), 'the matters that doctors embrace have far more serious importance than the matters with which the geometers occupy themselves.'

Though Louis had accrued disciples of his own, including influential Americans, his heritage might have faded had it not been for the publication in 1840 of probably the first text in medical statistics by Jules Gavarret, the Principes Généraux de Statistique Médicale. Gavaret had received mathematical training when entering the Ecole Polytechnique in 1829 to become an artillery officer, but resigned to become a physician, ultimately working with a colleague of Louis's, Gabriel Andral. He was familiar with the work of Simeon Denis Poisson and derived the likely 'limits of oscillation' for Louis' observations. His application of some of Poisson's formulae led to the calculation of what amounted to 99% confidence limits for the typhoid fever death rate observed by Louis. Quoted originally as 52/140 (37%), Gavaret demonstrated that the results could vary between 26% and 49%, though his conception of these limits would probably not have strictly paralleled that of a modern frequentist statistician. On one hand, Louis' followers were stung by the implied criticism, while the traditionalists still needed convincing. Even the great Claude Bernard was sceptical: 'By destroying the biological character of phenomena, the use of averages in physiology and medicine usually gives only apparent accuracy to the results. In a word, if based on statistics, medicine can never be anything but a conjectural science.'

There is some irony in the modern day epithet evidence-based medicine. The 9th edition of the Concise Oxford Dictionary describes evidence as (i) the available facts, circumstances, etc. supporting a belief (my italics) or proposition or indicating whether or not a thing is true or valid; (ii) statements or proofs admissible in a court of law. Poisson, in fact, following a tradition established by Laplace and the Marquis de Condorcet, was developing concepts of probability for their explicit application in a judicial process. Given certain characteristics of the accused and of the jury and the witness testimony, he attempted to derive the probability of reaching a correct verdict. This was at a time when the classical or Enlightenment view of probability could variously have been defined as: (i) degrees of certainty or probative weight; (ii) degrees of propensity; (iii) expectation (or practical probability in a game of chance); (iv) absence of purpose or design.

Although pre-dating a formal theory of statistical hypothesis testing by over 150 years, Daniel Bernoulli had used the latter definition in 1735 as a form of null hypothesis in calculating the probability that the inclinations of the five major planets were all less than seven degrees and thirty minutes. The same lines of reasoning had been used in 1752 by Maupertius in deriving the likelihood that polydactyly would be transmitted in three generations of the same family by chance.

For roughly a hundred years, therefore, theoretical developments would not have impeded the adoption of quantitative thinking in medicine and the development and application of 'limits of oscillation' by Gavarret was in essence a first attempt to surround clinical observations with confidence limits reflecting the variability incumbent on sample size. Enthusiastic German commentators welcomed Gavarret's contribution, urging their colleagues to 'avoid the illusions of praxis' but a very telling comment of Louis' foretells a lesson that the profession has been rather slower to learn: '... we shall hear no more of medical tact, of a kind of divining power of physicians.'

The limits of confidence

A burgeoning literature is now casting new light on how one's perspective and faulty heuristics can easily become pitfalls for sound clinical inference and decision making. Experts in a variety of fields usually make intuitive judgements that employ pattern recognition and a variety of mental rules of thumb. Sackett's view that clinical expertise (in judging whether external evidence applies to an individual patient) can be a sort of safety net disregards the possibility that assignment of diagnosis and subsequent prognostication might be influenced by both the choice of treatment options and the available evidence. The process by which new data are absorbed and formulated into heuristics is poorly understood, but there is ample evidence that 'expert' clinicians and novices alike are prone to a variety of judgement biases.

While the 'science' of evidence-based practice is now de rigueur, few medical curricula explicitly deal
with the shortcomings for clinical reasoning of a variety of cognitive and judgmental biases. Even specialists are prone to the vagaries of availability and representativeness heuristics, as well as base rate, primacy and anchoring effects. Feinstein, for example, demonstrated a form of ‘context bias’ when a group of radiologists blindly read the same pulmonary arteriograms twice within two distinct series. He showed clearly how the sequence of presentation and base-rate of abnormality affected their detection sensitivity.

In an out-patient based study, Bushyhead showed convincingly how test ordering behaviour can impose a ‘biased’ patterning of physical signs. Essentially a form of verification bias, the doctors in their study who ordered chest X-rays more frequently when they detected rales ended up with a biased estimate of the likelihood ratio for rales as a correlate of pneumonia. This probably contributes to the rather low correlation between the confidence in and accuracy of clinical prediction, irrespective of experience. Consultants often express surprise at this, but the reasons are not so mysterious: they seldom get comprehensive feedback on positive and negative outcomes (with the consequent delusions of verification bias); past predictions are often mistakenly recalled as over-consistent with actual outcomes; and, in hospital at least, they are usually exposed to skewed non-representative populations.

Some of the mis-match between clinical and actuarial prediction stems from poor judgement of probabilities. But at least one controlled trial has shown that teaching doctors to make better probability judgements may not always alter their treatment decisions. Perhaps we should be more ready to acknowledge that it is our beliefs as much as the evidence, which actually guides our actions. Freedman and Spiegelhalter showed in the early 1980s, when designing a trial of thiotepa for bladder cancer, how the minimum important clinical difference (or benefit) demanded of the new drug by 18 urologists varied by several fold among the 18 specialists. This is in keeping with the very variable medical decision making.

Mark Twain once wryly observed ‘Get your facts right first, then you can distort them as much as you please’. In clinical practice, the same facts expressed as relative or absolute risk reductions have been shown time and again to dictate different courses of action. As the bile in the exchanges above suggests, the facts never speak for themselves. The present day debate between Frequentists and Bayesians, and the fact that most ‘specialists’ are rather poor probability assessors to start with may serve to remind us that clinical expertise alone could be a rather threadbare safety net for Evidence-Based Medicine.

References


