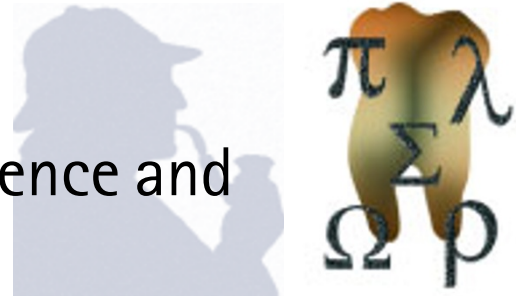


## IN BRIEF

- An insight into evidence-based dentistry and the origins of evidence-based medicine
- The philosophy of drawing conclusions from evidence
- Different approaches to assessing the evidence provided by the results in a frequency table
- An explanation of the number needed to treat (NNT)
- Some guidelines to follow when adopting an evidence-based approach to dentistry

## Further statistics in dentistry Part 10: Sherlock Holmes, evidence and evidence-based dentistry



J. F. Osborn<sup>1</sup> J. S. Bulman<sup>2</sup> and A. Petrie<sup>3</sup>

If one were to go by the explosion of interest in evidence-based clinical practice in the past decade of the second millennium, one could be forgiven for thinking that the idea was new. In fact, a quick search of *Medline* revealed 9,306 references to 'evidence-based medicine' (EBM) and 291 when the search was restricted to dentistry. It is claimed (Sackett *et al.*, 1996)<sup>1</sup> that the origins of EBM date back to mid nineteenth century Paris or earlier although the name EBM was coined in 1992. The inventor of the randomised controlled clinical trial, Sir Austin Bradford Hill, in the 1950s set out the statistical foundations of EBM.

### FURTHER STATISTICS IN DENTISTRY:

1. Research designs 1
2. Research designs 2
3. Clinical trials 1
4. Clinical trials 2
5. Diagnostic tests for oral conditions
6. Multiple linear regression
7. Repeated measures
8. Systematic reviews and meta-analyses
9. Bayesian statistics
10. Sherlock Holmes, evidence and evidence-based dentistry

It is not the intention of this article to review either the 9,306 articles or the 291 articles or even the substantial contributions made to evidence-based medicine published in this Journal. Rather, the objectives of this article are much more general:

1. To describe briefly what is evidence-based medicine and dentistry.
2. To describe the nature of external evidence. To review, very briefly, the philosophy of drawing conclusions from evidence.
3. To describe how evidence can be quantified.

### EVIDENCE-BASED MEDICINE (EBM)

In an important editorial (Sackett *et al.*, 1996) entitled, '*Evidence-Based Medicine: What it is and what it isn't*',<sup>1</sup> David Sackett, one of the pioneers of the new movement for the practice of EBM, and his colleagues emphasise that EBM (and by analogy, evidence-based dentistry) has two components '*The practice of EBM means integrating individual clinical expertise with the best available external clinical evidence from systematic research.*' Individual clinical expertise is acquired as a result of clinical practice and means that a clinician is not expected to slavishly follow rules dictated by others when it comes to the treatment of a particular patient. The clinician is likely to know much more about the needs of an individual patient, about the history of the condition, about the social context of the patient including his way of life, his family

background, employment situation etc than can be found by reading and learning from research reports, whose main objective is to reach generalised conclusions about 'patients of this type'. As Sherlock Holmes said, '*There is nothing like first hand evidence*', (Arthur Conan Doyle (ACD): *A Study in Scarlet*, 1888).<sup>2,3</sup> On the other hand, the results of excellent relevant clinical research provide a scientifically valid framework for patient care. According to Sackett *et al.* (1996),<sup>1</sup> '*External clinical evidence both invalidates previously accepted diagnostic tests and treatments and replaces them with new ones that are more powerful, more accurate, more efficacious and safer.*' Or in the words of Holmes, '*The mystery gradually clears away as each new discovery furnishes a step which leads to the complete truth*', (ACD *The Engineer's Thumb*, 1892). Clearly both components are necessary; clinical expertise without the application of the results of new research is likely to stagnate and cannot be expected to progress without the continuing education provided by good clinical publications. '*Education never ends, Watson. It is a series of lessons with the greatest for the last.*' (ACD *The Adventure of the Red Circle*, 1911).

### THE NATURE OF EXTERNAL EVIDENCE

Evidence is the ultimate product of the analysis of a series of observations. Such a statement may appear banal, but in fact, precise observations are a necessary ingredient for the improvement of clinical expertise and the production of good

<sup>1</sup>Professor of Epidemiological Methods, University of Rome, La Sapienza;

<sup>2</sup>Honorary Reader in Dental Public Health, Eastman Dental Institute for Oral Health Care Sciences, University College London;

<sup>3</sup>Senior Lecturer in Statistics, Eastman Dental Institute for Oral Health Care Sciences, University College London

\*Correspondence to: Aviva Petrie, Senior Lecturer in Statistics, Biostatistics Unit, Eastman Dental Institute for Oral Health Care Sciences, University College London, 256 Gray's Inn Road, London WC1X 8LD E-mail: a.petrie@eastman.ucl.ac.uk

### Refereed Paper

© British Dental Journal 2003; 194: 189–195

## Evidence

The weight of evidence derived from a clinical study will depend on its design and how well it has been conducted



research. There is a great tendency for all of us to observe what we expect to see rather than what actually occurs. Sometimes this problem can be ameliorated in clinical research by blinding the patient and the clinical observer and yes, even the statistician. Sir Arthur Conan Doyle was a medical practitioner and it is said that he modelled his fictional detective, Sherlock Holmes, on one of his professors at Edinburgh Medical School, Dr Joseph Bell (1837-1911). Bell was thought by his students to be a magician. In Doyle's words, 'Dr Bell would sit in a receiving room, with a face like a Red Indian, and diagnose people as they came in, before they even opened their mouths. He would tell them their symptoms and even give them details of their past life and hardly ever would make a mistake'.

Although the philosophical origins of EBM date back to the mid-nineteenth century, its legal status in Britain was implied by the Apothecaries Act of 1815, which licensed apothecaries in order to protect the public from the growing number of unqualified druggists and herbalists. The Medical Act of 1858 led to the creation of the medical register which contained the names of all doctors with recognised medical qualifications. The 1858 Act restricted the practice of medicine to those doctors included in the register. There was also the implication that these doctors should practice 'real' medicine, that is, the medicine taught and learned in medical schools, and the public would be protected against charlatans. The Act was not successful in eliminating complementary or alternative medicine, and indeed, apart from a short period in the middle of the twentieth century, the number of people who seek medical help outside the official medical profession, particularly from herbalists, has continued to increase. (Paradoxically, alternative medicine is still promoted and supplied by chemist's shops, the very place where a patient, having consulted a regular doctor, is required to go to collect his prescription! Boots, the chemists, even publish and distribute free a booklet (Anon, 2000)<sup>4</sup> in which complementary medicine is stated to be safe, and it is implied that orthodox medical help need be sought only where symptoms are severe and persistent). The 1858 Act implied that the medicine practised by registered doctors was based on evidence while the alternative was based on hearsay, old-wives-tales, grannies' remedies etc. If this distinction was one of the objectives of the 1858 Act, it most certainly was not very successful; there are many examples, in all medical specialities, of practice which, either for lack of evidence or ignorance, is not based on evidence. For example, Mills *et al.* (1994) showed that a much publicised and widely sold pre-brushing mouthwash was ineffective in reducing plaque or stains in comparison with a control,<sup>5</sup> while Scherer *et al.* (1998) showed that a herbal mouth rinse significantly reduced gingival bleeding.<sup>6</sup> Some controversies seem never to be resolved because of the difficulty of obtaining sufficient clear-cut evidence one way or the other. Is the use of mercury

amalgams totally without risk? Even if the dental profession is convinced of its safety, there are many that would not seem to be. Should pathology free impacted third molars be extracted prophylactically? Bandolier<sup>7</sup> answers by asking, 'What do you do when there is no evidence? Carry on with what you are doing because you have no evidence to stop, or stop what you are doing because there is no evidence to carry on?'. Similarly, Alexander (1998) has identified eleven myths of dentoalveolar surgery, and so on.<sup>8</sup>

The weight of the evidence derived from a clinical study will depend on its design and how well it has been conducted. A simple case series reporting a new treatment may not provide very strong evidence of the effect of the treatment unless the observed effect is exceedingly different from the natural progress of the disease or condition. On the other hand, a case series may be sufficient to generate a hypothesis which might be investigated by more rigorous studies. A control group will always increase the validity of a study based on a case series.

As noted in Part 3 of this series, Clinical Trials I, randomisation of the patients to the treatment groups will tend to remove the effect of confounding factors especially if the trial is not too small. Thus in terms of a single study, the randomised controlled trial (RCT) provides the best evidence that a treatment has an effect in comparison with the control group. This evidence is usually presented in the form of a significance test and a confidence interval for the treatment effect.

When there are several studies of the effect of a particular treatment, the results may be aggregated using the techniques of meta-analysis (Part 8 of this series), another new name for an old idea. 'There is nothing new under the sun, it has all been done before.' (ACD. *A Study in Scarlet*, 1888). To learn of the pitfalls of combining evidence in a meta-analysis, there is no better starting point than the early article by Daniels and Bradford Hill (1952).<sup>9</sup> A good meta-analysis should take account of the study designs, involve a well defined strategy for literature searches, assessment of quality, inclusion and exclusion criteria, tests of homogeneity etc, although in 1991, Thompson and Pocock (1991) felt that it was necessary to pose the question: can meta-analysis be trusted?<sup>10</sup> In the true spirit of meta-analysis, Holmes pleads, 'Any truth is better than infinite doubt' (ACD. *The Yellow Face*, 1893).

A review should bring together all the evidence for and against the effectiveness of a treatment, and there may be no simple clear-cut result. Further, there may be more than one review, and what should be done if the reviews differ in their conclusions? On the question of the extraction of impacted third molars, Bandolier suggests that the quality of the reviews be judged, and if they do not contain randomised controlled trials, they should be regarded with 'a cold and fishy eye', which leaves us just about where we started by asking the question: what



should the clinician do, stop or carry on? However, for a given patient, the clinician must make a decision and may not have the luxury Holmes enjoyed when he said in honesty to Watson, 'No, no; I never guess. It is a shocking habit - destructive to the logical faculty' (ACD. *The Sign of the Four*, 1890).

If the reviews do agree, a review of the reviews may evolve into a clinical guideline. One might be forgiven for thinking that at this point there would no longer be controversy, but not so. Whether created locally or nationally or internationally, guidelines are generally an aggregation of research evidence, expert opinion and clinical experience. The existence of a clinical guideline may intentionally have the effect of limiting the freedom of action of a clinician in the treatment of his patient, and this could have legal consequences and ethical implications. Holmes is mistaken when he says of Dr Grimesby Roylott 'When a doctor goes wrong, he is the first of criminals. He has nerve and the knowledge.' (ACD: *The Speckled Band* 1892). Holmes is speaking of going wrong in a legal sense rather than making a mistaken clinical judgement, but unfortunately a clinician rarely has all the knowledge, and errors will occur. Hurwitz (1998) expounds a comprehensive and highly readable account of the possible legal implications of following or not following clinical guidelines in his book appropriately titled *Clinical Guidelines and the Law; Negligence, Discretion and Judgement*.<sup>11</sup> These implications are important because the existence of guidelines neither implies that they will be followed in practice nor that their effectiveness will be formally evaluated. Not surprisingly, after meta-analyses of meta-analyses and reviews of reviews, there is also a *Guide to Guidelines* (Smith, 1997)!<sup>12</sup>

## EVIDENCE AND THE PHILOSOPHY OF SCIENTIFIC PROGRESS

The above title of this section is nothing but presumptuous when one thinks of the miles of shelves of books and other publications on this subject produced over the past 100 years, but statistics has played an important, under-rated and often overlooked role in the theories propounded by professional philosophers. Healy (2000) has recently published an entertaining but serious discussion of the role of statistics in the philosophy of science, and the philosophy of science in the practice of statistics.<sup>13</sup> In essence, the modern subject, 'statistics', has its origins at University College London around the start of the twentieth century when Karl Pearson began studying the theory of distributions and applying statistical methods to study biological problems, and, for example, discovered the chi-squared distribution and began thinking in terms of the significance test. Pearson's ideas were expanded and developed in the 1920s and 1930s by Fisher, and Gossett (the ever famous 'student' who first described the *t*-test)<sup>14</sup> perfected the idea of the statistical significance test

which has remained with us, virtually without change, until today. Basically the logical procedure followed in a statistical significance test is:

1. A stimulus provokes the need to perform an experiment to compare the effects of say, two treatments, A and B, on an outcome. The origin and form of the stimulus is not important and may come from a clinical observation, a hunch, hearsay, complimentary medicine etc. If the stimulus is based on evidence, this evidence cannot be used further in the experiment, and the experiment to compare the two treatments will be interpreted with a completely open mind, ignoring all that is known before (unless a Bayesian approach, discussed in Part 9, is used).
2. A null hypothesis is formulated, which states that the treatment effect (eg that the average difference between the two treatments) is zero. This hypothesis represents the state of knowledge at the start of the experiment and relates to the population of values.
3. The results of the experiment, derived from sample data, are analysed to discover if they provide sufficient evidence to reject the null hypothesis and thus change the state of knowledge by concluding that one treatment is better than the other. 'It is a capital mistake to theorise before one has the data. Insensibly one begins to twist the facts to suit the theories, instead of the theories to suit the facts.' (ACD. *A Scandal in Bohemia*, 1892). The decision to reject the null hypothesis, however, is based on probabilistic reasoning. Actually, it is the frequency or repeated experiment approach to probability as opposed to subjective probability or *a priori* probability reasoning. A single patient cannot of him or herself disprove the null hypothesis. 'We balance probabilities and choose the most likely. It is the scientific use of the imagination.' (ACD. *The Hound of the Baskervilles*, 1901-2).
4. A confidence interval for the effect of interest, such as the average difference between the treatments, is constructed. This should enable the researcher to determine whether or not there is sufficient evidence to conclude that the difference between the treatments is of *clinical* importance.

In fact, it was some decades later that the most influential philosopher of science of the twentieth century, Sir Karl Popper (1959 and 1972),<sup>15,16</sup> re-proposed that science advances, a step at a time, by the refutation of hypotheses. It seems that the statisticians were expert Popperians long before Popper's theories became popular. Fisher (1937), anticipating Popper by almost twenty years, stated 'Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.' and this assertion was based on a complex form of probabilistic reasoning.<sup>14</sup>

Later, Popper's theory was challenged by



### The statistical significance test

The statistical significance test is a procedure for deciding whether the evidence provided by the sample data is consistent with the null hypothesis about the population



Thomas Kuhn (1966), who argued that while science generally progresses slowly and steadily, there were events of dramatic importance, or revolutions, which totally changed the state of knowledge.<sup>17</sup> One can easily think of examples of such revolutions which have transformed scientific thinking: the introduction into Europe of the decimal number system by Leonardo di Pisa (Fibonacci) in 1202, enabling complex arithmetic to be performed and providing the trigger for the start of the renaissance, Galileo, Newton, Einstein's theory of relativity etc. However, revolutions also occur within specialities and in dentistry one can think of examples, such as the discovery of effective anaesthetics, the invention of the first high speed flexible shaft dental drill by Samuel Stockton White in 1844, the adoption of mercury based amalgams for fillings instead of gold about 160 years ago, or the discovery of the role of fluoride in the prevention of childhood caries. The introduction of randomisation in clinical trials by Bradford Hill was a revolution in medical statistics,<sup>18</sup> and it seems that maybe we are experiencing now a Kuhnian revolution in the form of the Bayesian approach to evidence from clinical studies. If indeed the subject, 'statistics', is transformed totally by the adoption of Bayesian techniques, it will be necessary to re-think what is meant by evidence based on Popperian inference in relation to medical practice. However, for the near or medium future, the validity and strength of evidence will continue to be based on Fisher-Popper statistical significance tests and their associated confidence intervals. Perhaps sadly, we are likely to witness for many more years the spectacle of our normally calm, serious, reserved research worker, suddenly triumphant and exuberant as his computer prints out the long awaited and much desired  $P < 0.05$ .

**Comparison by the difference between the two success rates**

Suppose the proportions of patients whose treatment result in a success are  $p_1 = a/n_1$  for the new and  $p_2 = b/n_2$  for the standard treatment. (These sample proportions are often referred to as the estimated success rates even if they are not strictly rates; in the population, the true success 'rates' are  $\pi_1$  and  $\pi_2$ ). The null hypothesis is that the two treatments have an equal chance of success (ie  $\pi_1 = \pi_2$ ). The statistical significance of the difference between  $p_1$  and  $p_2$  can be determined by calculating  $z_1 = (p_1 - p_2)/SE(p_1 - p_2)$  which follows the standardized Normal distribution, or equivalently, by calculating  $z_1^2$  which follows the chi-squared distribution with one degree of freedom [note:  $SE(p_1 - p_2)$  is the standard error of  $(p_1 - p_2)$ ]. A continuity correction should be applied to  $z_1$  and to the chi-squared test statistic but it has relatively little effect if the sample sizes are not too small. The 95% confidence interval for the true difference in the two success rates is  $(p_1 - p_2) \pm 1.96SE(p_1 - p_2)$ . These expressions for the significance test and the confidence interval assume that the values of  $a$ ,  $b$ ,  $c$  and  $d$  are not too small but if that were the case, Fisher's exact significance test would be an appropriate alternative to  $z_1$  or the chi-squared test statistic.

**Comparison by the difference between the two failure rates**

This is the half-empty version of the half-full glass. If the observed proportion of patients whose treatment fails is  $q_1$  for the new and  $q_2$  for the standard treatment, the difference between them is the same as the difference between the two success rates. Thus the significance of the difference between the two failure rates and the confidence interval for the difference are identical to those of the difference in success rates.

**Comparison by the ratio of the two success rates**

If the ratio of the two success rates is  $R_1 = p_1/p_2$ , the sampling distribution of  $R_1$  is log-Normal. Then the hypothesis that the ratio of the true success rates is one can be tested by calculating  $z_2 = \log_e R_1 / SE(\log_e R_1) = \log_e R_1 / \sqrt{(q_1/a + q_2/b)}$  which follows the standardized Normal distribution. It should be noted that  $z_2$  is not exactly equal to  $z_1$ . The 95% confidence interval for the ratio of the true success rates is obtained by calculating the exponential of the two limits for  $\log_e$  of this ratio, ie the exponential of  $\log_e R_1 \pm 1.96SE(\log_e R_1)$ .



**The Bayesian approach**

It will be necessary to rethink what is meant by evidence based on Popperian or classical inference if the Bayesian approach to statistical analysis is adopted

**PRESENTING STATISTICAL EVIDENCE**

We all know that there is not much difference between a half-full glass and one that is half-empty, but the results of even the simplest research can be presented in a bewildering variety of ways. Consider the simplest clinical trial in which a new treatment is to be compared with the existing standard treatment and the outcome is dichotomous, a success or a failure. The results would usually be set out in the form of Table 1, a two by two table of frequencies. The number of ways of comparing the two treatments and their associated tests of significance and confidence intervals seem limitless.

**Table 1 2x2 table of frequencies**

	New treatment	Standard treatment	Total	
Success	$a$	$b$	$a+b$	total number of successes
Failure	$c$	$d$	$c+d$	total number of failures
Total	$n_1$	$n_2$	$n$	total number of patients



**Comparison by the ratio of the two failure rates**

If the ratio of the two failure rates is  $R_2 = q_2/q_1$ , the Standard Normal Deviate for testing the significance of the ratio of the true failure rates is  $z_3 = \log_e R_2 / SE(\log_e R_2) = \log_e R_2 / \sqrt{(p_1/c + p_2/d)}$  which is not exactly equal to  $z_1$  or  $z_2$ . The 95% confidence interval for the ratio of the true failure rates is obtained by calculating the exponential of the two limits for  $\log_e$  of this ratio, ie the exponential of  $\log_e R_2 \pm 1.96SE(\log_e R_2)$ .

**Comparison by the odds ratio of success and the odds ratio of failure**

The observed odds of a success for the new treatment is  $a/c$  and for the standard treatment it is  $b/d$ . The observed odds ratio for a success is thus  $OR_1 = (a/c)/(b/d) = (ad)/(bc)$ ; the Standard Normal Deviate for testing the significance of the true odds ratio for a success is:

$$z_4 = \log_e OR_1 / SE(\log_e OR_1) = \log_e OR_1 / \sqrt{(1/a + 1/b + 1/c + 1/d)}$$

A 95% confidence interval for the true odds ratio is the exponential of the limits for  $\log_e$  of this ratio, ie the exponential of

$$\log_e OR_1 \pm 1.96\sqrt{(1/a + 1/b + 1/c + 1/d)}$$

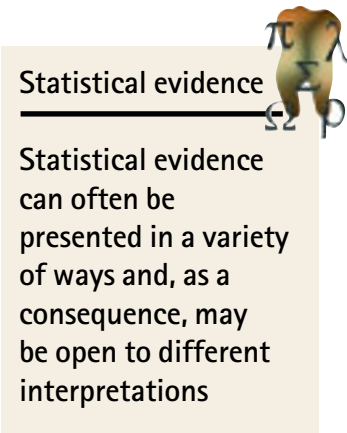
The value of  $z_4$  is not exactly equal to  $z_1$ ,  $z_2$  or  $z_3$ . If instead of the odds of a success, the odds of failure are considered,  $OR_2 = (bc)/(ad)$  which is the reciprocal of  $OR_1$ . The Standard Normal Deviate for the test of significance of the true odds ratio of a failure is  $-z_4$  and the 95% confi-

dence interval is the reciprocal of that for the odds ratio of a success. The odds ratio gives equivalent results from the significance tests for both success and failure.

**Comparison of the different methods**

These four methods are probably the most commonly used of the very many methods of summarising a 2x2 table of frequencies. Table 2 shows the bewildering array of results obtained from applying these methods to three simple examples. In each of the three examples the new treatment doubles the success rate of the old treatment. However, it can be seen that the method of comparison may give different impressions of the improvement offered by the new treatment even if the methods only produce slightly different Standard Normal Deviates for the significance test of the null hypothesis, and slightly different P-values.

In the first example, the success rates for the standard and the new treatments are both low (ie 0.05 and 0.10) and hence the difference between them is small. The new treatment doubles the success rate but the old method only increases the failure rate by 5.6% (ie  $5/90 \times 100\%$ )! The Fisher exact test and the corrected Standard Normal Deviate (or, equivalently, the corrected chi-squared test) give the same P-values, which are different from those obtained using the other methods. This is because the 2x2 table contains small frequencies (the numbers of patients with a successful outcome are 5 and 10) and the corrected and Fisher exact values are probably the most reliable, the other P-values being too small



Statistical evidence can often be presented in a variety of ways and, as a consequence, may be open to different interpretations

**Table 2. Results of the analyses of three examples of hypothetical clinical trials.**

Observed frequencies	Example 1		Example 2		Example 3			
	new	standard	new	standard	new	standard		
Success	a	b	10	5	20	10	80	40
Failure	c	d	90	95	80	90	20	60
<i>Estimated effects</i>								
diff. $p_1 - p_2$	0.10-0.05=0.05		0.20-0.10=0.10		0.80-0.40=0.40			
diff. $q_2 - q_1$	0.95-0.90=0.05		0.90-0.80=0.10		0.60-0.20=0.40			
ratio $p_1/p_2$	2.000		2.000		2.000			
ratio $q_2/q_1$	1.056		1.125		3.000			
OR for success (new/standard)	2.111		2.250		6.000			
<i>Test statistics</i>								
$z_1$ (difference)	1.34		1.98		5.77			
$z_1$ (corrected)	1.23		1.78		5.63			
$z_2$ (ratio $p_1/p_2$ )	1.31		1.92		5.24			
$z_3$ (ratio $q_2/q_1$ )	1.34		1.96		5.09			
$z_4$ (odds ratio)	1.32		1.95		5.55			
<i>P-values</i>								
$P_1$ (difference)	0.180		0.048		<0.000001			
$P_1$ (corrected)	0.283		0.075		<0.000001			
$P_1$ (Fisher exact)	0.283		0.073		<0.000001			
$P_2$ (ratio $p_1/p_2$ )	0.190		0.055		<0.000001			
$P_3$ (ratio $q_2/q_1$ )	0.182		0.050		<0.000001			
$P_4$ (odds ratio)	0.188		0.051		<0.000001			



## Number needed to treat (NNT)



The NNT is the number of patients that need to have the new treatment instead of the old in order to have one additional patient benefit or, equivalently, prevent one adverse reaction

because of the lack of the continuity correction.

The second example is even more perplexing. The ratio of the two success rates is two and the odds ratio is 2.25 but using the old treatment only increases the risk of failure by 12.5%! If a decision were to be made on the basis of the 5% level of significance there would be even greater difficulty, since for some comparisons  $P \leq 0.05$  while for others  $P > 0.05$ ! Again in this example, the Fisher test and the corrected Standard Normal Deviate are probably the most reliable and the conclusion should be that the difference is not statistically significant at the 5% level.

In the third example, the success rates are comparatively large and there are no very low frequencies in the  $2 \times 2$  table. The new treatment doubles the chance of success from 40% to 80% and increases the odds of success 6-fold whilst the use of the old treatment triples the risk of failure. The values of  $z$ , although slightly different, lead to exactly the same interpretation, that it is most unlikely that the observed difference between the treatments is due merely to chance.

The discussion of these three examples has concentrated on statistical significance only because the evaluation of evidence is to a large degree based on the Fisher-Popperian philosophy that only by the refutation of hypotheses can scientific knowledge progress. It is left as an exercise for the reader to calculate the associated confidence intervals, which would be essential if the usefulness of a real new treatment were to be evaluated in comparison with the old. Although the data in the three examples are hypothetical, it may be disturbing that even with objective statistical analysis, the results may be open to different interpretations. It is not just a case of a glass being half full or half empty. As Holmes observed *'There is nothing more deceptive than an obvious fact'* (ACD *The Boscombe Valley Mystery*, 1891).

### The number needed to treat (NNT)

Because it can be difficult to interpret differences between treatments, other methods of comparing treatments that have a more direct clinical interpretation have been investigated. One of these, the *'number needed to treat'* (NNT) is becoming increasingly popular. McQuay and Moore (1997) give a full description of the method,<sup>19</sup> but the basic idea is to calculate the number of patients that need to have the new treatment instead of the old in order to have one additional patient benefit or, equivalently, prevent one adverse outcome. In the third of the examples discussed earlier, the success rates are 0.80 and 0.40 in the new and standard treatment groups, respectively. The  $2 \times 2$  table shows that if 100 patients have the new treatment instead of the old, there will be 40 more successes. Thus to achieve just one more success, it will be necessary to give  $100/40 = 2.5$  patients the new treatment. It is not difficult to see that the NNT is just the reciprocal of the difference between the two success rates.

Thus  $NNT = 1/(p_1 - p_2) = 1/(0.8 - 0.4) = 1/0.40 = 2.5$ , or equivalently  $NNT = 1/(q_2 - q_1)$ .

Since the standard error of  $(p_1 - p_2)$  in this example is  $\sqrt{[p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2]} = 0.0632$ , the 95% confidence interval for the true difference is  $0.40 \pm 1.96 \times 0.0632$ , or 0.276 to 0.524. Thus the 95% confidence interval for the true NNT is 1.9 to 3.6 (ie  $1/0.524$  to  $1/0.276$ ). The values of NNT to achieve one more success for the first and second examples are 20 and 10 patients, respectively.

The possible values for NNT range from one, when the old treatment is useless and the new is perfect, to infinity when there is no difference between the treatments. Not only does it express the result of the comparison between the treatments in terms of a number of patients, and thus a concept more readily grasped by a clinician than, for example, an odds ratio, but also it may have direct application in the cost/benefit analysis of the decision to adopt the new treatment. This is not to say that the other measures are never useful: indeed each one may be appropriate for a given situation. In an example from dentistry, discussed in Part 8 of this series, a review (Rijkom *et al.*, 1997) of the usefulness of fluoride gel for the prevention of caries, showed that the overall effect of the gel is to reduce the incidence of caries by 22% per year.<sup>20</sup> Calculations showed that, if without the gel the incidence of caries were 0.25 DMFS per year, it would be necessary to treat 18 teeth with the gel for a year in order to save one DMFS. The NNT decreased to 9, 4.5 and 3 if the incidences without the gel were 0.50, 1.00 and 1.50 DMFS per year, respectively. This example is particularly interesting because it demonstrates that where caries is more prevalent, the NNT is less; in other words, where caries is rare, it may not be worthwhile to use the gel, but where it is common, it might be very cost effective indeed.

### SOME GUIDELINES

So, recognising the many problems facing the clinician, how can he/she use the evidence-based approach to greatest effect? According to the guidelines propounded by Sackett *et al.* (2000), the following sequence of steps, incorporating the statistical principles described in earlier papers in this series, should be pursued:<sup>21</sup>

**Step 1** Convert the need for information about prevention, diagnosis, prognosis, therapy, etc, into an answerable question which relates specifically to the patient's requirements and the population of interest.

**Step 2** Track down the best evidence with which to answer that question using, for example, MEDLINE and evidence databases (such as Evidence-based Medicine Reviews from Ovid Technologies ([www.ovid.com](http://www.ovid.com)) which combines several electronic databases including the Cochrane Database of Systematic Reviews).

**Step 3** Critically appraise the evidence for its validity (closeness to the truth), impact (size of the effect), and applicability (usefulness in clinical practice).



cal practice). This involves ensuring that sources of potential bias have been eliminated, that the appropriate statistical methods have been used, and that all the important outcomes have been considered and are appropriately summarised (eg rates, NNT) with confidence intervals so that a decision can be made as to whether or not the results are clinically important.

**Step 4** Integrate the critical appraisal with clinical expertise and with the patient's unique biology, values and circumstances.

**Step 5** Finally, evaluate performance in terms of effectiveness and efficiency by questioning the ability to complete steps 1-4 successfully, and seek ways to improve performance in the future.

## CONCLUSION

It is well known that the number of research journals and research papers increases at an alarming rate every year. It would be hoped that the growth in the number of good research reports is equally rapid. If this is in fact the case, in future it will be ever more difficult to identify good research and maintain a register of valid evidence. The Cochrane Foundation, the National Health Service Centre for Reviews and Dissemination at the University of York and others have taken an enormous step forward by trying to filter out the valid evidence from the bulk of less worthy research. Certainly individual clinicians cannot be expected to read all the latest research reports in their field, let alone evaluate them and classify the results as good evidence or not. *I consider that a man's brain is like a little empty attic, and you have to stock it with such furniture as you choose. A fool takes in all the lumber of every sort that he comes across, so that the knowledge which might be useful to him gets crowded out, or at best is jumbled up with a lot of other things, so that he has a difficulty in laying his hands upon it. Now the skilful workman is very careful indeed as to what he takes into his brain-attic. He will have nothing but the tools which may help him in doing his work, but of these he has a large assortment, and all in the most perfect order. It is a mistake to think that that little room has elastic walls and can distend to any extent. Depend upon it – there comes a time when for every addition of knowledge you forget something you knew before. It is of the highest importance, therefore, not to have useless facts elbowing out the useful ones.* (ACD *A Study in Scarlet*, 1887). A clinician, therefore, either must become ever more specialised and remember only the very important aspects of his narrow field, or he can remain a general practitioner but he has to accept that in many situations he will have to consult his 'library'. Nowadays, it is almost essential to have a computer to keep the lumber-room in an accessible order and enable easy contact to be made with such

organisations as the Cochrane Foundation and the NHS Centre for Reviews and Dissemination.

Sir Arthur Conan Doyle and his mentor, Dr Joseph Bell, were acutely aware of the value of good evidence in medical practice and for the detective work of Sherlock Holmes. The 60 stories involving Holmes and his assistant Dr Watson were published between 1887 and 1927 in the *Strand Magazine*, *Colliers Weekly* and other periodicals. They have given pleasure to generations of avid readers eager to discover something of the extraordinary ability of Sherlock Holmes to deduce the truth from whatever evidence was available. Pearson, Fisher, 'Student', Popper and others have formalised the idea of the use of evidence to test hypotheses and enable science to progress. Bell, Sackett, his colleagues and others have sought to identify from the mass of available research evidence what is valid and can be realistically applied in the every day practice of clinical medicine.

- 1 Sackett D L, Rosenberg W M C, Gray J A M, Haynes R B, Richardson W S. Evidence-based medicine: what it is and what it isn't. *Br Med J* 1996; **312**: 71-72.
- 2 All the Sherlock Holmes stories may be found in Arthur Conan Doyle (1981) or at [www.sherlockian.net/canon/index.html](http://www.sherlockian.net/canon/index.html)
- 3 Arthur Conan Doyle (1981) *The Penguin Complete Sherlock Holmes*. Penguin
- 4 Anon (2000) *Your guide to choosing holistic medicines*. Complementary Medicine The Boots Company PLC, Nottingham, England.
- 5 Mills D C, Smith S R, Chung L. The effect of using a pre-brushing mouthwash (Plax) on removal of tooth stain in vivo and in vitro. *J Clin Periodont* 1994; **21**: 13-16.
- 6 Scherer W, Gultz J, Sangwoo Lee S, Kaim J. The ability of a herbal mouthrinse to reduce gingival bleeding. *J Clin Dent* 1998; **9**: 97-100.
- 7 Bandolier. Prophylactic removal of impacted third molars. <http://www.jr2.ox.ac.uk/Bandolier/band42/b42-2.html>
- 8 Alexander R E. Eleven myths of dentoalveolar surgery. *J Am Dent Assoc* Sep 1998; **129**: 1271-1279.
- 9 Daniels M, Hill A B. Chemotherapy of pulmonary tuberculosis in young adults; an analysis of the combined results of three Medical Research Council Trials. *Br Med J* 1952; **i**:1162-1168.
- 10 Thompson S G, Pocock S. Can meta-analysis be trusted? *Lancet* 1991; **338**: 1127-1130.
- 11 Hurwitz B. *Clinical Guidelines and the Law, Negligence, Discretion and Judgement*. Radcliffe Medical Press, Abingdon, 1998.
- 12 Smith P. (ed) *Guide to the Guidelines: Disease management made simple*. Abingdon: Radcliffe Medical Press, 1997.
- 13 Healy M J R. Paradigms and Pragmatism: Approaches to Medical Statistics. *Ann Ig* 2000; **12**.
- 14 Fisher R A. Prof Karl Pearson and the method of moments. *Ann Eugenics* 1937; **7**: 303-318.
- 15 Popper K. *Conjectures and Refutations*. 4th ed., London: Routledge and Kegan Paul, 1972.
- 16 Popper K. *The Logic of Scientific Discovery*. London: Hutchinson, 1959.
- 17 Kuhn T. *The structure of Scientific revolutions*. 3rd ed. The University of Chicago Press, 1966.
- 18 Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. *Br Med J* 1948; **769**-782.
- 19 McQuay H J, Moore R A. Using numerical results from systematic reviews in clinical practice. *Ann Int Med* 1997; **126**: 712-720.
- 20 Van Rijkom H M, Truin G J, Van't Hof M A. A meta-analysis of clinical studies on the caries-inhibiting effect of fluoride gel treatment. *Caries Res* 1998; **32**: 83-92.
- 21 Sackett D L, Strauss S E, Richardson W S, Rosenberg W, Haynes R B. *Evidence-based Medicine: How to Practice and Teach EBM*. 2nd Edn. Churchill Livingstone, 2000.



Cochrane  
Collaboration

The Cochrane Collaboration is an international organization that helps people make well informed decisions about healthcare by preparing, maintaining and ensuring the accessibility of systematic reviews of the effects of health care interventions