Part 1

Assessing and utilizing the diagnostic or prognostic power of biomarkers

1

Evidence-based medicine: evaluation of biomarkers

R Andrew Moore

Pain Research and Nuffield Department of Anaesthetics, University of Oxford, Oxford, UK

Evidence-based medicine

Evidence-based medicine (EBM) has been described as the '*conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients*' [1]. Because there are so many biomedical journals (perhaps as many as 30 000), the chance of any practitioner being aware of all the developments of interest is vanishingly small. The philosophy of EBM, therefore, extends into ways of summarizing information to make it understandable and useful. The key tool is the systematic review, and most work on systematic reviews, and indeed on EBM, has concentrated on treating disease.

Systematic review

Reviews are called systematic when they include a thorough search for all published (and sometimes unpublished) information on a topic. Empirical observation in systematic reviews of treatment efficacy demonstrates several sources of bias occurring because of the architecture of study design. The ones we know of are:

Randomization	Nonrandomized studies can overestimate treatment effects by up to 40%, or even change the conclusions of a review. Including only randomized studies is likely to be sensible for reviews of the effectiveness of treatments.
Blinding	Open (nonblinded) studies overestimate treatment effects by about 17%.
Quality	Studies of lower reporting quality overestimate treatment effects.
Quantity	Small studies can overestimate treatment effects.

Duplication	Trials may be reported more than once. This may be legitimate, but is often incorrect and without cross-referencing. Unrecognized duplicate publications can lead to an overestimation in treatment effects of 20%.

Now, not all of these sources of bias will occur in each circumstance, but some will, and there may be others that are yet to be identified. What the systematic review process teaches us about trials of effectiveness is that there are many sources of potential bias, and we may not know all of them. It is notable is that every one we know of tends to overestimate the effects of treatment. There are other factors that may be important as potential sources of bias, particularly issues relating to the validity of experimental design in specific clinical situations.

Since systematic reviews concentrate on all the worthwhile published material on a topic, they provide the basis for a fresh look at where we are. One of their main results is to refresh the research agenda. A particular example is the increasing concentration on outcomes – the change in a disease state that is worthwhile for patients, their carers or the healthcare system. All too often, research concentrates on what is measurable, rather than what is meaningful. The large, simple, clinical trial with patient-defined outcomes may be one of the most important developments of EBM.

Size

Clinical trials are performed in order to tell whether one treatment is better than another. The statistical power of the trial is calculated on the basis of being able to say with confidence that there is a difference. It is the direction of the effect that is being measured. However, most of the time what we really want to know is the magnitude of the effect of treatment. To do this, we need much more information – perhaps 10 times as many patients need to be studied.

Figure 1.1 shows the results of 56 meta-analyses of placebo in about 12000 patients in acute pain trials [2]. Overall, 18% of patients given placebo had more than 50% pain relief over 6 hours. All trials in the meta-analyses were randomized, all were double blind and all had the same outcomes measured in the same way. The variability with small samples is huge, from 0% to nearly 60%. Only when the sample is above 1000 patients given placebo is the true rate measured.

This is just one example of how small studies can be affected by random chance. This should not be surprising: calculating confidence intervals around small samples will demonstrate that uncertainty is large with small samples. However, it serves to illustrate the power of random variation with the use of small samples, and why it is dangerous to extrapolate from a single small trial to clinical practice.

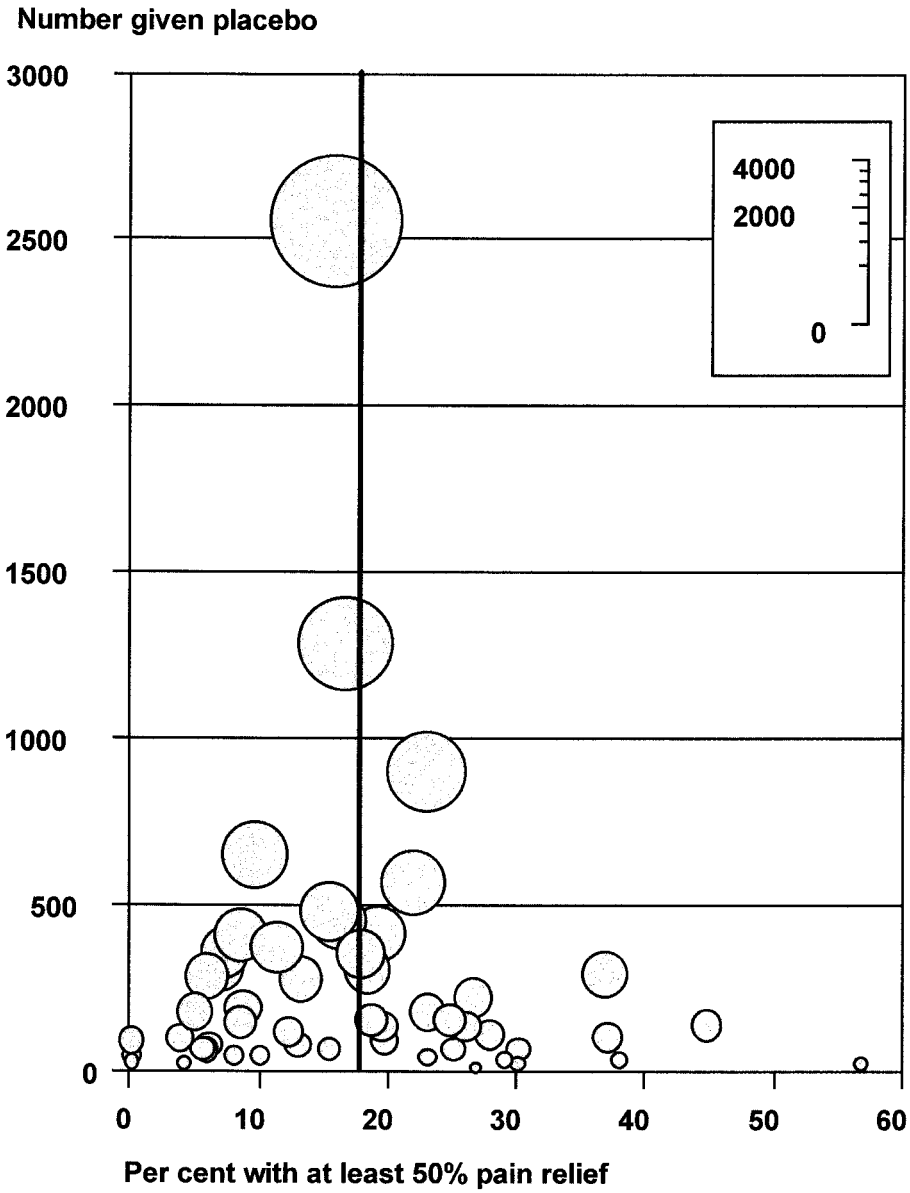

Figure 1.1 Per cent of patients with at least 50% pain relief from meta-analyses of acute pain studies. Each symbol represents one meta-analysis; all trials were randomized and double blind and with the same outcome measured over the same time (2). Size of the symbol is proportional to the number of patients included. The vertical line shows the overall average response (18%) from over 12 000 placebo patients.

Expressing results

EBM has a real problem in how to express the results of research so that they can be understood and used. Statistical significance is in itself an unhelpful output, as are odds ratios, risk ratios, relative risks, weighted mean differences or effect sizes. The simple fact is that few people understand them and even fewer can use them.

What catapulted EBM into the real world was the use of the number-needed-to-treat (NNT). This is the inverse of the absolute risk reduction and describes the therapeutic effort required to produce one patient with the required clinical outcome [3]. It has proved particularly useful when there are many different treatments, as in analgesics for pain. By producing tables of NNTs for analgesics, choice can be made in terms of efficacy, harm and cost.

However, better understanding of the requirements of large samples to assess clinical outcomes accurately [4] is likely to lead to even simpler outcomes than the NNT. The future holds the prospect of being able to say, with confidence, that a given treatment in patients with a given disease and severity will lead to a successful outcome in x% – which would be understandable by doctors, patients and policy makers.

Evidence-based laboratory medicine

There are various types of evidence we accept for laboratory tests and biomarkers: evidence about the analytical performance of an assay; evidence about quality control in the laboratory and quality assurance from external schemes; and evidence about issues like sensitivity and specificity in particular clinical circumstances. What we rarely have, though, is evidence that the use of a laboratory test can, for a given patient or group of patients, make a clinically relevant difference to the diagnosis or treatment. Evidence-based laboratory medicine (EBLM) has to encompass all of these types of evidence, of course, but the judgement will increasingly be made on clinical outcomes.

Whether systematic reviews will be helpful for EBLM, as they have been for treatments, is questionable, however. One description of levels of evidence commonly used for studies of diagnostic tests is shown in Table 1.1. The keys to good quality have been said to be independence, masked comparison with a reference standard and consecutive patients from an appropriate population. Lower quality comes from inappropriate populations and comparisons that are not masked or with different reference standards. Until recently, we lacked any empirical or theoretical evidence about the levels of bias that any of these study architectures can impart.

A new contribution from Holland [5] provides the missing link. The authors searched for and found 26 systematic reviews of diagnostic tests with at least five included studies. Only 11 could be used in their analysis, because 15 were either not

Table 1.1. Levels of evidence for studies of diagnostic methods

Level	Criteria
1	An independent, masked comparison with reference standard among an appropriate population of consecutive patients
2	An independent, masked comparison with reference standard among nonconsecutive patients or confined to a narrow population of study patients
3	An independent, masked comparison with an appropriate population of patients, but reference standard not applied to all study patients
4	Reference standard not applied independently or masked
5	Expert opinion with no explicit critical appraisal, based on physiology, bench research or first principles

systematic in their searching or did not report any sensitivity or specificity. Data from the remainder were subjected to mathematical analysis, to investigate whether the presence or absence of some item of proposed study quality made a difference to the perceived value of the test.

There were 218 individual studies, only 15 of which satisfied all eight criteria of quality that this analysis concerned. Thirty per cent fulfilled at least six of eight criteria. To evaluate bias, the authors calculated the relative diagnostic odds ratio by comparing the diagnostic performance of a test in those studies that failed to satisfy the methodological criterion with the performance of the test in studies that did meet this criterion. Overestimation of effectiveness (positive bias) of a diagnostic test was shown by a lower confidence interval for the relative diagnostic odds ratio of more than 1.

The results are shown in Table 1.2. Use of different reference tests, lack of blinding and lack of a description of either the test or the population in which the test was studied led to positive bias. However, the largest factor leading to positive bias was evaluation of a test in a group of patients already known to have the disease and a separate group of normal patients – called a case-control study in the paper [5].

There are also pointers to good practice in the publication of articles on diagnostic tests. The authors of a most important paper [6] set out seven methodological standards (Table 1.3). They then looked at papers published in the *Lancet*, *British Medical Journal*, *New England Journal of Medicine* and *Journal of the American Medical Association* from 1978 through 1993 to see how many reports of diagnostic tests meet these standards. Between 1978 and 1993, they found 112 articles, predominantly on radiological tests and immunoassays. Few of the standards were met consistently – ranging from 51% avoiding workup bias down to 9% reporting accuracy in subgroups (Table 1.3). While there was an overall improvement over

Table 1.2. Empirical evidence of bias in diagnostic test studies of different architecture

Study characteristic	Relative diagnostic odds ratio (95% CI)	Description
Case-control	3.0 (2.0–4.5)	A group of patients already known to have the disease compared with a separate group of normal subjects
Different reference tests	2.2 (1.5–3.3)	Different reference tests used for patients with and without the disease
Not blinded	1.3 (1.0–1.9)	Interpretation of test and reference is not blinded to outcomes
No description of test	1.7 (1.1–1.7)	Test not properly described
No description of population	1.4 (1.1–1.7)	Population under investigation not properly described
No description of reference	0.7 (0.6–0.9)	Reference standard not properly described

Note:
The relative diagnostic odds ratio indicates the diagnostic performance of a test in studies failing to satisfy the methodological criterion relative to its performance in studies with the corresponding feature [5].

time for reports to score on more standards, even in the most recent period studied only 24% met up to four standards, and only 6% up to six.

Most diagnostic test evaluations are structured to examine patients with a disease compared with those without the disease – a case-control design. Astonishingly, few studies are performed according to the highest standard in Table 1.1. The studies which have been published are seriously flawed, as Read et al. [6] have demonstrated. It must be questioned, therefore, whether any systematic review of diagnostic tests is worthwhile.

Size

Just as large samples are needed to overcome the random effects of chance for treatments, so they are also needed for tests. An example is the controversy over falling sperm counts. A meta-analysis [7] collected 61 studies on sperm counts published between 1938 and 1990. Almost one-half of these studies (29/61) studied fewer than 50 men. The smallest number was nine and the largest 4435 men. Only 2% of the data on nearly 15000 men was collected before 1970, in small studies. Figure 1.2 shows the variability by size. The overall mean sperm count was 77 million/ml, but small individual studies recorded means from 40 to 140 million/ml. Only large studies correctly estimated the overall mean, and any temporal relationship is spurious because the old studies were small.

Table 1.3. Standards of reporting quality for studies of diagnostic tests

Reporting standard	Background	Criteria	Per cent meeting standard
Spectrum composition	The sensitivity and specificity of a test depend on the characteristics of the population studied. Change the population and you change these indices. Since most diagnostic tests are evaluated on populations with more severe disease, the reported values for sensitivity and specificity may not be applicable to other populations with less severe disease in which the test will be used.	For this standard to be met, the report had to contain information on any three of these four criteria: age distribution, sex distribution, summary of presenting clinical symptoms and/or disease stage, and eligibility criteria for study subjects.	27
Pertinent subgroups	Sensitivity and specificity may represent average values for a population. Unless the condition for which a test is to be used is narrowly defined, then the indices may vary in different medical subgroups. For successful use of the test, separate indices of accuracy are needed for pertinent individual subgroups within the spectrum of tested patients.	This standard is met when results for indices of accuracy were reported for any pertinent demographic or clinical subgroup (for example, symptomatic versus asymptomatic patients).	9
Avoidance of workup bias	This form of bias can occur when patients with positive or negative diagnostic test results are preferentially referred to receive verification of diagnosis by the gold standard procedure.	For this standard to be met in cohort studies, all subjects had to be assigned to receive both the diagnostic test and the gold standard verification either by direct procedure or by clinical follow up. In case-control studies, credit depended on whether the diagnostic test preceded or followed the gold standard procedure. If it preceded, credit was given if disease verification was obtained for a consecutive series of study subjects regardless of their diagnostic test result. If the diagnostic test followed, credit was given if test results were stratified according to the clinical factors which evoked the gold standard procedure.	51

Table 1.3. (*cont.*)

Reporting standard	Background	Criteria	Per cent meeting standard
Avoidance of review bias	This form of bias can be introduced if the diagnostic test or the gold standard is appraised without precautions to achieve objectivity in their sequential interpretation – like blinding in clinical trials of a treatment. It can be avoided if the test and gold standard are interpreted separately by persons unaware of the results of the other.	For this standard to be met in either prospective cohort studies or case-control studies, a statement was required regarding the independent evaluation of the two tests.	43
Precision of results for test accuracy	The reliability of sensitivity and specificity depends on how many patients have been evaluated. Like many other measures, the point estimate should have confidence intervals around it, which are easily calculated.	For this standard to be met, confidence intervals or standard errors must be quoted, regardless of magnitude.	12
Presentation of indeterminate test results	Not all tests come out with a black or white, yes/no, answer. Sometimes they are equivocal, or indeterminate. The frequency of indeterminate results will limit a test's applicability, or make it cost more because further diagnostic procedures are needed. The frequency of indeterminate results and how they are used in calculations of test performance represent critically important information about the test's clinical effectiveness.	For this standard to be met, a study had to report all of the appropriate positive, negative or indeterminate results generated during the evaluation and whether indeterminate results had been included or excluded when indices of accuracy were calculated.	26
Test reproducibility	Tests may not always give the same result – for a whole variety of reasons of test variability or observer interpretation. The reasons for this, and its extent, should be investigated.	For this standard to be met in tests requiring observer interpretation, at least some of the tests should have been evaluated for a summary measure of observer variability. For tests without observer interpretation, credit was given for a summary measure of instrument variability.	26

Source: From Read et al. [6].

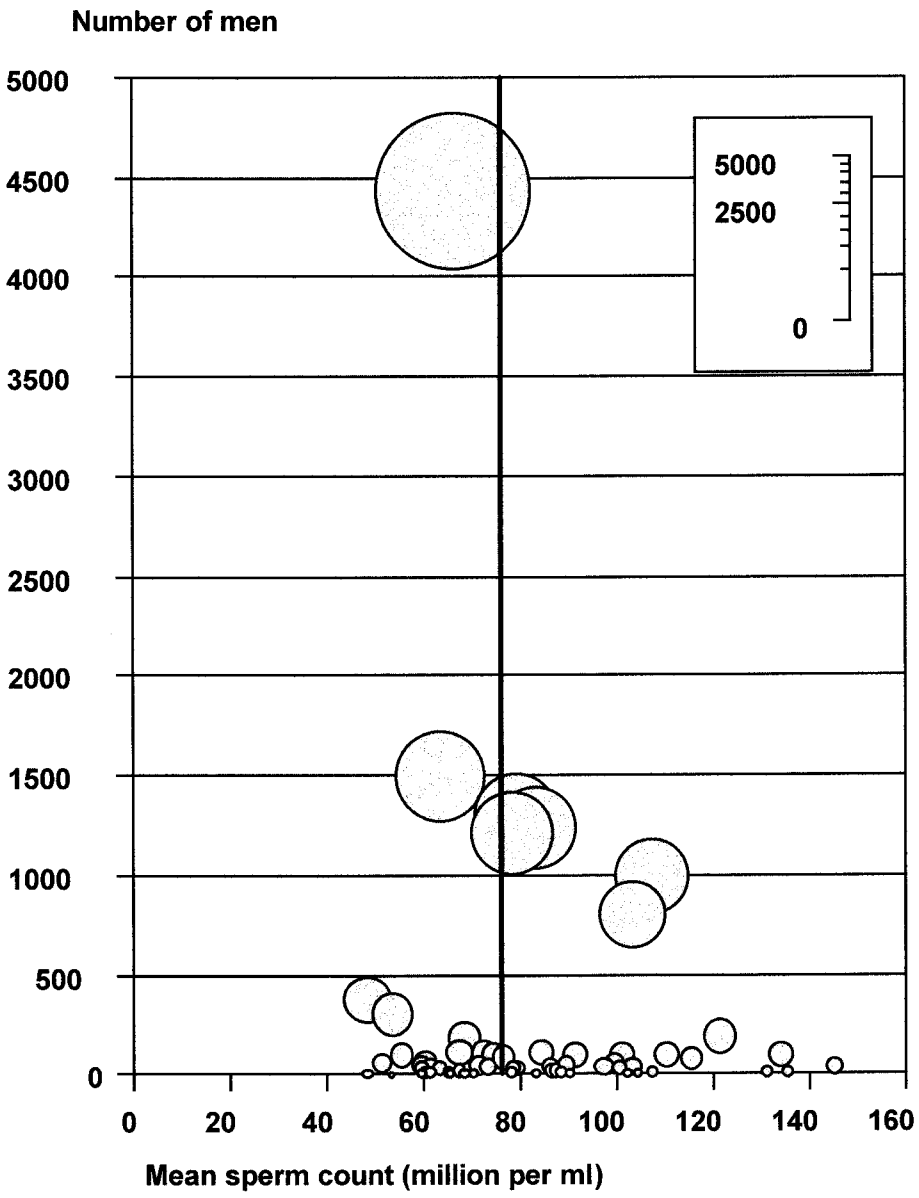

Figure 1.2 Mean sperm counts from individual studies in a meta-analysis (7). Each symbol represents one study. Size of the symbol is proportional to the number of patients included. The vertical line shows the overall mean (77 million/ml) from over 15 000 men.