PART 1
Understanding evidence and pain

CHAPTER 1
Why evidence matters

Andrew Moore and Sheena Derry
Pain Research, Nuffield Department of Anaesthetics, John Radcliffe Hospital, Oxford, UK

Introduction

There are two ways of answering a question about what evidence-based medicine (EBM) is good for or even what it is. One is the dry, formal approach, essentially statistical, essentially justifying a proscriptive approach to medicine. We have chosen, instead, a freer approach, emphasizing the utility of knowing when "stuff" is likely to be wrong and being able to spot those places where, as the old maps would tell us, "here be monsters." This is the *Bandolier* approach, the product of the hard knocks of a couple of decades or more of trying to understand evidence.

What both of us (and Henry McQuay and other collaborators over the years), on our different journeys, have brought to the examination of evidence is a healthy dose of skepticism, perhaps epitomized in the birth of *Bandolier*. It came during a lecture on evidence-based medicine by a public health doctor, who proclaimed that only seven things were *known* to work in medicine. By *known*, he meant that they were evidenced by systematic review and meta-analysis. A reasonable point, but there were unreasonable people in the audience. One mentioned thiopentone for induction of anesthesia, explaining that with a syringe and needle anyone, without exception, could be put to sleep given enough of this useful barbiturate; today we would say that it had an NNT of 1. So now we had seven things known to work in medicine, plus thiopentone. We needed somewhere to put the bullet points of evidence; you put bullets in a bandolier (a shoulder belt with loops for ammunition).

The point of this tale is not to traduce well-meaning public health docs, or meta-analyses, but rather to make the point that evidence comes in different ways and that different types of evidence have different weight in different circumstances. There is no single answer to what is needed, and we have often to think outside what is a very large box. Too often, EBM seems to be corralled into a very small box, with the lid nailed tightly shut and no outside thinking allowed.

If there is a single unifying theory behind EBM, it is that, whatever sort of evidence you are looking at, you need to apply the criteria of quality, validity, and size. These issues have been explored in depth for clinical trials, observational studies, adverse events, diagnosis, and health economics [1], and will not be rehearsed in detail in what follows. Rather, we will try to explore some issues that we think are commonly overlooked in discussions about EBM.

We talk to many people about EBM and those not actively engaged in research in the area are frequently frustrated by what they see as an impossibly complicated discipline. Someone once quoted Ed Murrow at us, who, talking about the Vietnam war, said that "Anyone who isn't confused doesn't really understand the situation" (Walter Bryan, *The Improbable Irish*, 1969). We understand the sense of confusion that can arise, but there are good reasons for continuing to grapple with EBM. The first of these is all about the propensity of research and other papers you read to be wrong. You need to know about that, if you know nothing else.

Evidence-Based Chronic Pain Management. Edited by C. Stannard, E. Kalso and J. Ballantyne.

Most published research false?

It has been said that only 1% of articles in scientific journals are scientifically sound [2]. Whatever the exact percentage, a paper from Greece [3], replete with Greek mathematical symbols and philosophy, makes a number of important points which are useful to think of as a series of little laws (some of which we explore more fully later) to use when considering evidence.

- The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
- The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.
- The greater the number and the fewer the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
- The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.
- The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true. (These might include research grants or the promise of future research grants.)
- The hotter a scientific field (the more scientific teams involved), the less likely the research findings are to be true.

Ioannidis then performs a pile of calculations and simulations and demonstrates the likelihood of us getting at the truth from different typical study types (Table 1.1). This ranges from odds of 2:1 on (67% likely to be true) from a systematic review of good-quality randomized trials, through 1:3 against (25% likely to be true) from a systematic review of small inconclusive randomized trials, to even lower levels for other study architectures.

There are many traps and pitfalls to negotiate when assessing evidence, and it is all too easy to be misled by an apparently perfect study that later turns out to be wrong or by a meta-analysis with impeccable credentials that seems to be trying to pull the wool over our eyes. Often, early outstanding results are followed by others that are less impressive. It is almost as if there is a law that states that first results are always spectacular and subsequent ones are mediocre: the law of initial results. It now seems that there may be some truth in this.

Three major general medical journals (*New England Journal of Medicine*, *JAMA*, and *Lancet*) were searched for studies with more than 1000 citations published between 1990 and 2003 [4]. This is an extraordinarily high number of citations when you think that most papers are cited once if at all, and that a citation of more than a few hundred times is almost as rare as hens' teeth.

Of the 115 articles published, 49 were eligible for the study because they were reports of original clinical research (like tamoxifen for breast cancer prevention or stent versus balloon angioplasty). Studies had sample sizes as low as nine and as high as 87,000. There were two case series, four cohort studies, and 43 randomized trials. The randomized trials were very varied in size, though, from 146 to 29,133 subjects (median 1817). Fourteen of the 43 randomized trials (33%) had fewer than 1000 patients and 25 (58%) had fewer than 2500 patients.

Of the 49 studies, seven were contradicted by later research. These seven contradicted studies included

Table 1.1 Likelihood of truth of research findings from various typical study architectures

Example	Ratio of true to not true
Confirmatory meta-analysis of good-quality RCTs	2:1
Adequately powered RCT with little bias and 1:1 prestudy odds	1:1
Meta-analysis of small, inconclusive studies	1:3
Underpowered and poorly performed phase I–II RCT	1:5
Underpowered but well-performed phase I–II RCT	1:5
Adequately powered exploratory epidemiologic study	1:10
Underpowered exploratory epidemiologic study	1:10
Discovery-orientated exploratory research with massive testing	1:1000

one case series with nine patients, three cohort studies with 40,000–80,000 patients, and three randomized trials, with 200, 875 and 2002 patients respectively. So only three of 43 randomized trials were contradicted (7%), compared with half the case series and three-quarters of the cohort studies.

A further seven studies found effects stronger than subsequent research. One of these was a cohort study with 800 patients. The other six were randomized trials, four with fewer than 1000 patients and two with about 1500 patients.

Most of the observational studies had been contradicted, or subsequent research had shown substantially smaller effects, but most randomized studies had results that had not been challenged. Of the nine randomized trials that were challenged, six had fewer than 1000 patients, and all had fewer than 2003 patients. Of 23 randomized trials with 2002 patients or fewer, nine were contradicted or challenged. None of the 20 randomized studies with more than 2003 patients were challenged.

There is much more in these fascinating papers, but it is more detailed and more complex without becoming necessarily much easier to understand. There is nothing that contradicts what we already know, namely that if we accept evidence of poor quality, without validity or where there are few events or numbers of patients, we are likely, often highly likely, to be misled.

If we concentrate on evidence of high quality, which is valid, and with large numbers, that will hardly ever happen. As Ioannidis also comments, if instead of chasing some ephemeral statistical significance we concentrate our efforts where there is good prior evidence, our chances of getting the true result are better. This may be why clinical trials on pharmaceuticals are so often significant statistically, and in the direction of supporting a drug. Yet even in that very special circumstance, where so much treasure is expended, years of work with positive results can come to naught when the big trials are done and do not produce the expected answer.

Limitations

Whatever evidence we look at, there are likely to be limitations to it. After all, there are few circumstances in which one study, of whatever architecture, is likely to be able to answer all the questions we need to know about an intervention. For example, trials capturing information about the benefits of treatment will not be able to speak to the question of rare, but serious, adverse events.

There are many more potential limitations. Studies may not be properly conducted or reported according to recognized standards, like CONSORT for randomized trials (www.consort-statement.org), QUOROM for systematic reviews, and other standards for other studies. They may not measure outcomes that are useful, or be conducted on patients like ours, or present results in ways that we can easily comprehend; trials may have few events, when not much happens, but make much of not much, as it were. Observational studies, diagnostic studies, and health economic studies all have their own particular set of limitations, as well as the more pervasive sins of significance chasing, or finding evidence to support only preconceptions or *idées fixes*.

Perfection in terms of the overall quality and extent of evidence is never going to happen in a single study, if only because the ultimate question – whether this intervention will work in this patient and produce no adverse effects – cannot be answered. The average results we obtain from trials are difficult to extrapolate to individuals, and especially the patients in front of us (of which more later).

Acknowledging limitations

Increasingly we have come to expect authors to make some comment about the limitations of their studies, even if it is only a nod in the direction of acknowledging that there are some. This is not easy, because there is an element of subjectivity about this. Authors may also believe, with some reason, that spending too much time rubbishing their own results will result in rejection by journals, and rejection is not appreciated by pointy-headed academics who live or die by publications.

Even so, the dearth of space given over to discussing the limitations of studies is worrying. A recent survey [5] that examined 400 papers from 2005 in the six most cited research journals and two open-access journals showed that only 17% used at least one word denoting limitations in the context of the scientific work presented. Among the 25 most cited journals, only one (*JAMA*) asks for a comments section on study limitations, and most were silent.

Statistical testing

It is an unspoken belief that to have a paper published, it helps to report some measure with a statistically significant difference. This leads to the phenomenon of significance chasing, in which data are analyzed to death and the aim is to find any test with any data that show significance at the paltry level of 5%. A P value of 0.05, or significance at the 5% level, tells us that there is a 1 in 20 chance that the results occurred by chance. As an aside, you might want to ask yourself how happy you are with 1 in 20; after all, if you throw two dice, double six seems to occur frequently and that is a chance of 1 in 36. If you want to examine evidence with a cold and fishy eye, try recognizing significance only when it is at the 1 in 100 level, or 1%, or a P value of 0.001; it often changes your view of things.

Multiple statistical testing

The perils of multiple statistical testing might have been drummed into us during our education but as researchers, we often forget them in the search for "results," especially when such testing confirms our pre-existing biases. A large and thorough examination of multiple statistical tests underscores the problems this can pose [6].

This was a population-based retrospective cohort study which used linked administrative databases covering 10.7 million residents of Ontario aged 18–100 years who were alive and had a birthday in the year 2000. Before any analyses, the database was split in two to provide both derivation and validation cohorts, each of about 5.3 million persons, so that associations found in one cohort could be confirmed in the other cohort.

The cohort comprised all admissions to Ontario hospitals classified as urgent (but not elective or planned) using DSM criteria, and ranked by frequency. This was used to determine which persons were admitted within the 365 days following their birthday in 2000, and the proportion admitted under each astrological sign. The astrological sign with the highest hospital admission rate was then tested statistically against the rate for all 11 other signs combined, using a significance level of 0.05. This was done until two statistically significant diagnoses were identified for each astrological sign.

In all, 223 diagnoses (accounting for 92% of all urgent admissions) were examined to find two statistically significant results for each astrological sign. Of these, 72 (32%) were statistically significant for at least one sign compared with all the others combined. The extremes were Scorpio, with two significant results, and Taurus, with 10, with significance levels of 0.0003 to 0.048.

The two most frequent diagnoses for each sign were used to select 24 significant associations in the derivation cohort. These included, for instance, intestinal obstructions and anemia for people with the astrological sign of Cancer, and head and neck symptoms and fracture of the humerus for Sagittarius. Levels of statistical significance ranged from 0.0006 to 0.048, and relative risk from 1.1 to 1.8 (Fig. 1.1), with most being modest.

Protection against spurious statistical significance from multiple comparisons was tested in several ways.

When the 24 associations were tested in the validation cohort, only two remained significant: gastrointestinal haemorrhage and Leo (relative risk 1.2), and fractured humerus for Sagittarius (relative risk 1.4).

Using a Bonferoni correction for 24 multiple comparisons would have set the level of significance acceptable as 0.002 rather than 0.05. In this case, nine of 24 comparisons would have been significant in the derivation cohort, but none in the both derivation and validation cohort. Correcting for all 14,718 comparisons used in the derivation cohort would

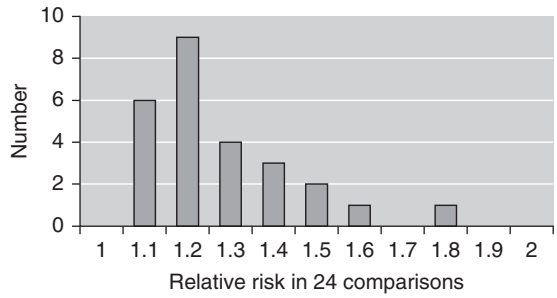

Figure 1.1 Relative risk of associations between astrological sign and illness for the 24 chosen associations, using a statistical significance of 0.05, uncorrected for multiple comparisons.

have meant using a significance level of 0.000003, and no comparison would have been significant in either derivation or validation cohort.

This study is a sobering reminder that statistical significance can mislead when we don't use statistics properly: don't blame statistics or the statisticians, blame our use of them. There is no biologic plausibility for a relationship between astrological sign and illness, yet many could be found in this huge data set when using standard levels of statistical significance without thinking about the problem of multiple comparisons. Even using a derivation and validation set did not offer complete protection against spurious results in enormous data sets.

Multiple subgroup analyses are common in published articles in our journals, usually without any adjustment for multiple testing. The authors examined 131 randomized trials published in top journals in 6 months in 2004. These had an average of five subgroup analyses, and 27 significance tests for efficacy and safety. The danger is that we may react to results that may have spurious statistical significance, especially when the size of the effect is not large.

Size is everything

The more important question, not asked anything like often enough, is whether any statistical testing is appropriate. Put another way, when can we be sure that we have enough information to be sure of the result, using the mathematical perspective of "sure," meaning the probability to a certain degree that we are not being mucked about by the random play of chance? This is not a trivial question, given that many results, especially concerning rare but serious harm, are driven by very few events.

In a clinical trial of drug A against placebo, the size of the trial is set according to how much better drug A is expected to be. For instance, if it is expected to be hugely better, the trial will be small but if the improvement is not expected to be large, the trial will have to be huge. Big effect, small trial; small effect, big trial; statisticians perform power calculations to determine the size of the trial beforehand. But remember that the only thing being tested here is whether the prior estimate of the expected treatment effect is actually met. If it is, great, but when you calculate the effect size from that trial, using number needed to treat (NNT), say, you probably have insufficient information to do so because the trial was never designed to measure the *size of the effect*. If it were, then many more patients would have been needed.

In practice, what is important is the size of the effect – how many patients benefit. With individual trials we can be misled. Figure 1.2 shows an example of six large trials (213–575 patients, 2000 in all) of a single oral dose of eletriptan 80 mg for acute migraine, using the outcome of headache relief (mild or no pain) at 2 hours. NNTs measured in the individual trials range from 1.6 to 3.1, an almost two-fold difference in the estimate of the size of the effect (overall, the NNT was 2.6). Even with these excellent trials, impeccably conducted, variations in response with eletriptan (between 56% and 69% in individual trials) and placebo (between 21% and 40%) mean that there is uncertainty over the size of the effect. For many treatments and dose/drug/condition combinations, we have much less information, fewer events, and much more uncertainty over the size of the effect.

Consider Figure 1.3, which looks at the variation in the response to placebo in over 50 meta-analyses in acute pain. In all the 12,000 or more patients given placebo, the response rate was 18% (meaning not that placebo **caused** 18% of people to have at least 50% pain relief over 6 hours, but that 18% of people in trials like

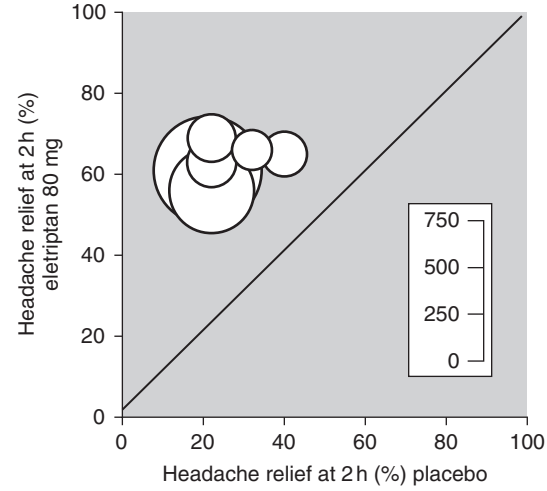

Figure 1.2 Headache response at 2 hours for oral eletriptan 80 mg. Size of symbol is proportional to number of patients in a trial.

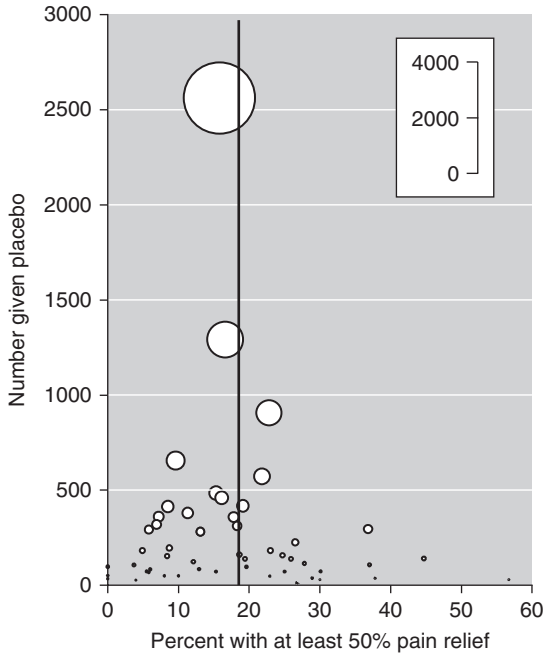

Figure 1.3 Percentage of patients with at least 50% pain relief with placebo in 56 meta-analyses in acute pain. Size of symbol is proportional to number of patients given placebo. Vertical line is the overall average.

these will have at least 50% pain relief over 6 hours if you do nothing at all). With small numbers, the measured effect with placebo varies from 0% to almost 50%. Only when the numbers are large is there greater consistency, and there are many other examples like this of size overcoming variability caused by the random play of chance.

How many events?

A few older papers keep being forgotten. When looking at the strengths and weaknesses of smaller meta-analyses versus larger randomized trials, a group from McMaster suggested that with fewer than 200 outcome events, research (meta-analyses in this case) may only be useful for summarizing information and generating hypotheses for future research [7]. A different approach using simulations of clinical trials and meta-analyses arrived at pretty much the same conclusion, that with fewer than 200 events, the magnitude and direction of an effect become increasingly uncertain [8].

Just how many events are needed to be reasonably sure of a result when event rates are low (as is the case for rare but serious adverse events) was explored some while ago [9]. This looks at a number of examples, varying event rates in experimental and control groups, using probability limits of 5% and 1%, and with lower and higher power to detect any difference. Higher power, greater stringency in probability values, lower event rates, and smaller differences in event rates between groups all suggest the need for more events and larger numbers of patients in trials. Once event rates fall to about 1% or so, and differences between experimental and control to less than 1%, the number of events needed approaches 100 and number of patients rises to tens of thousands.

All of which points to the inescapable conclusion that with few events, our ability to make sense of things is highly impaired. As a rule of thumb, we can probably dismiss studies with fewer than 20 events, be very cautious with 20–50 events, and reasonably confident with more than 200 events – if everything else is OK.

Subgroup analyses

Almost any paper you read, be it analysis of a clinical trial, an observational study or meta-analysis of either, will involve some form of subgroup analysis, such as severity of condition, age or sex. In addition to the problems of multiple testing, subgroup analyses also tend to involve small numbers – because the more you slice and dice the data, the fewer the number of actual events – and, if they are clinical trials, remove the benefits of randomization. They almost always introduce the danger of some unknown confounding.

One of the best examples of the dangers of subgroup analysis, due to unknown confounding, comes from a review article examining the 30-day outcome of death or myocardial infarction from a meta-analysis of platelet glycoprotein inhibitors [10]. Analysis indicated different results for women and men (Fig. 1.4), with benefits in men but not women. Statistically this was highly significant ($P<0.0001$).

In fact, it was found that men had higher levels of troponins (a marker of myocardial damage) than women and when this was taken into account, the difference between men and women was understandable, with more effect with greater myocardial damage; sex wasn't the source of the difference.

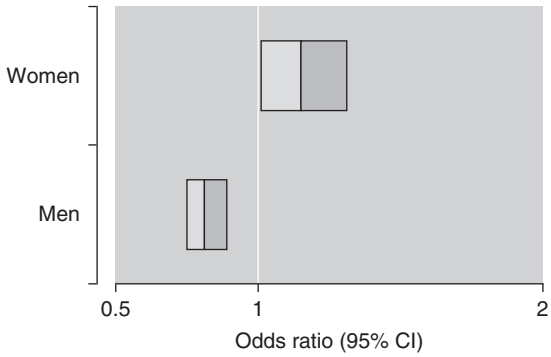

Figure 1.4 Subgroup analysis in women and men of death or MI with platelet glycoprotein inhibitors (95% confidence interval).

Trivial differences

It is worth remembering what relative risks tell us in terms of raw data (Table 1.2). Suppose we have a population in which 100 events occur with our control intervention, whatever that is. If we have 150 events with an experimental, the relative risk is now 1.5. It may be statistically significant, but most events were those occurring anyway. If there were 250 events, the relative risk would be 2.5, and now most events would occur because of the experimental intervention.

Large relative risks may be important, even with more limited data. Small relative risks, probably below 2.0 and certainly below about 1.5, should be treated with caution, especially where the number of

Table 1.2 Rules of causation

Feature	Comment
Consistency and unbiasedness of findings	Confirmation of the association by different investigators, in different populations, using different methods
Strength of association	Two aspects: the frequency with which the factor is found in the disease, and the frequency with which it occurs in the absence of the disease. The larger the relative risk, the more the hypothesis is strengthened
Temporal sequence	Obviously, exposure to the factor must occur before onset of the disease. In addition, if it is possible to show a temporal relationship, as between exposure to the factor in the population and frequency of the disease, the case is strengthened
"Biologic gradient (dose–response relationship)"	Finding a quantitative relationship between the factor and the frequency of the disease. The intensity or duration of exposure may be measured
Specificity	If the determinant being studied can be isolated from others and shown to produce changes in the incidence of the disease, e.g. if thyroid cancer can be shown to have a higher incidence specifically associated with fluoride, this is convincing evidence of causation
Coherence with biologic background and previous knowledge	The evidence must fit the facts that are thought to be related, e.g. the rising incidence of dental fluorosis and the rising consumption of fluoride are coherent
Biologic plausibility	The statistically significant association fits well with previously existing knowledge
Reasoning by analogy	Common sense, especially when you have other similar examples for types of intervention and outcome
Experimental evidence	This aspect focuses on what happens when the suspected offending agent is removed. Is there improvement? The evidence of remission – or even resolution of significant medical symptoms – following explanation obviously would strengthen the case It is unethical to do an experiment that exposes people to the risk of illness, but it is permissible and indeed desirable to conduct an experiment, i.e. a randomized controlled trial on control measures. If fluoride is suspected of causing thyroid dysfunction, for example, the experiment of eliminating or reducing occupational exposure to the toxin and conducting detailed endocrine tests on the workers could help to confirm or refute the suspicion

events is small, and even more especially outside the context of the randomized trial.

The importance of a relative risk of 2.0 has been accepted in US courts [11]. "A relative risk of 2.0 would permit an inference than an individual plaintiff's disease was more likely than not caused by the implicated agent. A substantial number of courts in a variety of toxic substance cases have accepted this reasoning."

Confounding by indication

Bias arises in observational studies when patients with the worst prognosis are allocated preferentially to a particular treatment. These patients are likely to be systematically different from those not treated or treated with something else (paracetamol rather than nonsteroidal anti-inflammatory drugs (NSAID) in asthma, for instance).

Confounding, by factors known or unknown, is potentially a big problem, because we do not know what we do not know and the unknown could have big effects, like troponin above. When relative risks are small, say below about 1.3, potential bias created because of unknown confounding, or confounding by indication improperly adjusted, becomes so great that it makes any conclusion at best unreliable. This is especially important when interpreting observational studies that appear to link a particular intervention with a particular outcome.

Adverse events

Evidence around adverse events is important, complicated, yet often poor. It is impossible to do justice to adverse event evidence in a few paragraphs, so perhaps it is worth sticking to the highlights.

Adverse events are important because the "value" of a particular therapeutic intervention depends on both potential benefit and potential harm in the individual. To assess this trade-off, we need evidence for both, and while evidence about benefit is generally well documented, at least in clinical trials of newer interventions, evidence about harm has been neglected.

Long-term drug therapy is increasingly being used for primary prevention. Asymptomatic patients may be asked to tolerate adverse effects when the likelihood of therapeutic benefit is small. Adverse events are a major influence on compliance and the most common reason for discontinuation in clinical practice. A medicine not taken is one that cannot work. There is an increasing tendency for more openness and accountability in clinical decision making, with patients asking for more information and taking a more active role in their care.

Adverse events occur in the absence of treatment, something to remember when looking at data. Symptoms commonly listed as adverse events in clinical trials happen to all of us at some time. Fortunately most of them are not serious and even if severe, are reversible. Most are not related to any therapeutic intervention. Groups of medical and nonmedical people in the USA in the 1960s [12], and medical students in Germany in the 1990s [13], who were free of disease and not in any kind of trial or taking any medication, were asked about symptoms. Most participants were in their 20s. They were given a list of symptoms and asked to record whether or not they had experienced any in the previous 3 days. Overall, 83% experienced at least one of the symptoms and only 17% reported none. There were no major differences between medical and nonmedical participants, or between studies carried out 30 years apart. The most common symptom reported by at least 40% was fatigue. Having an idea of the background rate of an adverse event in a study population is important as it can affect tolerability, and also how easy it is to establish a causal association with the intervention.

Another example of common adverse events would be constipation, something we worry about a lot when prescribing opioids. Constipation occurs in about 15% of people with chronic pain using weak opioids [14].

The overall average percentage of people with constipation in a systematic review of constipation prevalence in the US was about 15% (1 in 7 adults [15]). The range was 1.9–27%, depending to some extent on how constipation was ascertained. Most reports were in the range of 12–19%, with some self-reported prevalence being higher and two face-to-face questioning reports below 4%. There was a distinctly higher prevalence in women compared with men in almost every study, irrespective of method of ascertainment. Prevalence of constipation in women

was on average about twice as high as in men. There was also a consistent finding of higher constipation prevalence in non-Caucasian people, by a factor of about 1.4 to 1, though nonwhite racial groups were not subdivided. Other trends were for decreased prevalence in people with highest income and highest educational attainment or years of education, though these may well be measuring different aspects of the same phenomenon. Older age, especially age over 70 years, was also associated with higher constipation rates.

With any examination of adverse events, it is worth bearing in mind that what we want to establish is causation. The most important *aide-mémoire* is the Bradford-Hill rules, summarized in Table 1.2. They ask about strength of association, timing, dose–response, and other linking evidence. We need more than association to proceed to causation.

Safety

Claims are all too often made about safeties that are unfounded. To some extent, it depends what one means by safety, but members of the public say that they want to know about any adverse event that occurs at a rate more frequently than 1 in 100,000 [16]. To be even remotely confident about an adverse event occurring at a rate 10 times more frequently than that (1 in 10,000), we would need information from about 2 million people.

Clinical trials, even meta-analyses of clinical trials, will not have this amount of information. Nor will most observational studies or even meta-analyses of observational studies. Things may be changing, because large databases are beginning to be interrogated to provide data on safety. Caution is required because of confounding by indication and small numbers of events, so that individual studies can give very different results. For instance, a systematic review looking at NSAIDs and risk of myocardial infarction showed that the risk for naproxen compared to non-use of NSAIDs varied linearly between a relative risk of 0.5 and one of 1.5 (with a mean of 1.0).

Large database studies may also surprise. A good example of the surprising results of database studies (good as in good study, as well as a surprising result) indicated that long-term use of proton pump inhibitors significantly increased risk of hip fracture in older people [17]. It might be that the risk of using a proton pump inhibitor with NSAID incurs a bigger risk to life from hip fracture than did the gastrointestinal bleed the proton pump inhibitor was protecting against.

In any event, claims of absolute safety cannot be made, and we will see more examples of rare but serious adverse events in future than ever we did in the past.

Importance of the individual patient

The two quotations below come from people who argued vehemently over the role and importance of EBM yet agreed on the importance of the individual within the system.

> "Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patient." [18].

> "Managers and trialists may be happy for treatments to work on average; patients expect their doctors to do better than that." [19].

This underlines the importance of looking at information from the point of view of the individual patient. In acute pain, patients have been shown generally to obtain pain relief that is either very good or poor, but the average of responses to analgesics is at a point where there are few, if any, patients [20]. It is commonly understood that not every patient with a particular condition benefits from treatments known to work (on average). Patients may discontinue therapy because of adverse events as well as lack of efficacy, especially in chronic conditions. A clinical trial may tell us that 50% of patients have pain relief with drug, compared with 20% with placebo, and we applaud a good NNT of 3.3. Yet that obscures the fact that half the patients do not have pain relief but may have adverse effects.

A classic example demonstrating how different we all are is provided by a trial in which depressed patients were randomized to one of three antidepressants which were, on average, the same [21]. Patients initially randomized to one treatment frequently changed to another. By 9 months only 44% were still taking the treatment to which they had been randomized. Some (about 15%) were lost to follow-up after baseline or when on any of the randomized treatments. Others either switched to another antidepressant or

stopped treatment because of adverse effects or lack of efficacy, again without any difference between the three antidepressants. Each was taken by about the same proportion, on average, just different patients to those initially randomized. Patients and their doctors found the balance of effect and absence of adverse events that was right for them, and almost 70% had a good outcome over the 9 months of the trial.

The degree of variability between individuals in their physiologic response to drugs is remarkable, and best exemplified by a study of 50 healthy young volunteers who received rofecoxib 25 mg, celecoxib 200 mg or placebo in randomized order, and who underwent a series of tests [22]. There was considerable variability between individuals in cyclooxygenase 2 inhibition achieved, and in selectivity, for both of the drugs. Variation between individuals was 50 to several hundred-fold in activity in different *in vitro* tests following a single dose. Differences were associated with genetic polymorphisms and other factors were involved in the variability observed. Similarly, a range of polymorphisms in genes coding for enzymes metabolizing morphine, opioid receptors, and blood–brain barrier transport of morphine by drug receptors all contribute to considerable variability between individuals [23]. A number of mechanisms can influence individual responses to analgesics [24].

There are important practical implications following these findings. They obviously relate particularly to the potential harm of limited formularies, but also challenge how we use average results from trials in making decisions about individual patients.

Outcomes

Where evidence can often let us down is in the outcomes chosen in trials. Outcomes used may not be what we, or patients, want from treatment, but rather what it is possible to measure. Ideally, a satisfactory outcome should involve both benefit and lack of adverse events, because adverse events are often a cause of discontinuation of an otherwise effective therapy.

Things are changing. In migraine, for example, an outcome of mild or no pain 2 hours after therapy changed to no pain at 2 hours, then no pain at 2 hours plus no recurrence or need to use analgesics over the next 24 hours. The hurdle was getting higher. It was recently raised yet again, when an individual patient meta-analysis identified those patients who were both pain free for 24 hours and had no adverse effects [25]; this amounted to no more than 22% of the total, only 12% more than with placebo. A large randomized comparison of two triptans found about 30% of patients with this outcome [26].

There are other examples where people have sought more relevant outcomes. For instance, a series of different outcomes related to wart clearance and return emerged from a systematic review of genital wart therapy [27], while a longitudinal survey of patients with bipolar disorder suggested that success be judged over longer periods because of the sustained nature of the disorder [28].

There is no reason why we cannot demand more intelligent and comprehensive outcomes to be measured in clinical trials. While it is likely that the combination of benefit plus absence of adverse events will be found only in the minority, this will be a spur for both better use of what therapies we have and determination of better therapies for the future.

Conclusion

Evidence-based medicine is about a number of things. First and foremost, it is about avoiding being misled. That means that we have to have a passing acquaintance with issues of quality, validity, and size, and most of these come down to good old common sense. When a trial is done using two men and a dog and reports a subgroup analysis on the dog as statistically significant, that is not a reason for rushing to change practice.

The second thing that EBM should be about is making things better. This could mean wanting better and more meaningful outcomes or knowing how to assess trial results in terms of an individual patient, or asking the question of knowing which patient will benefit before you treat. It may be slow, but keeping some of these issues in your mind can mean hours of fun asking awkward questions of visiting speakers, a few of which may do some good. Over and above this, of course, is the incorporation of prior evidence in the production of new evidence, especially clinical trials, which are becoming bigger and better, though much more expensive to conduct.

Thirdly, when we collect together all the good evidence on a topic and get rid of the misleading, we often see more clearly. A number of examples exist in pain, especially in acute pain [29], migraine [30], and neuropathic pain [31].

The final message should be about the importance of wisdom. EBM, in its fullest sense, should incorporate evidence from whatever source, your knowledge of the patient, the patient's own preferences, and the circumstances you are in. Evidence should be regarded as a tool, not a rule. Even where there is limited evidence, in combination with clinical experience and wisdom it can produce useful results, perhaps the best example being a treatment algorithm for neuropathic pain [31].

References

1. Moore RA, McQuay HJ. *Bandolier's Little Book of Understanding the Medical Evidence*. Oxford University Press, Oxford, 2006.
2. Smith R. Where is the wisdom: the poverty of medical evidence. *BMJ* 1991; **303**: 798–799.
3. Ioannidis JPA. Why most published research findings are false. *PLoS Med* 2005; **2**: e124. www.plosmedicine.org.
4. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005; **294**: 218–228.
5. Ioannidis JPA. Limitations are not properly acknowledged in the scientific literature. *J Clin Epidemiol* 2007; **60**: 324–329.
6. Austin PC, Mamdani MM, Juurlink DN, Hux JE. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol* 2006; **59**: 964–969.
7. Flather MD, Farkouh ME, Pogue JM, Yusuf S. Strengths and limitations of meta-analysis: larger studies may be more reliable. *Control Clin Trials* 1997; **18**: 568–579.
8. Moore RA, Gavaghan D, Tramer MR, Collins SL, McQuay HJ. Size is everything – large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. *Pain* 1998; **78**: 209–216.
9. Shuster JJ. Fixing the number of events in large comparative trials with low event rates: a binomial approach. *Control Clin Trials* 1993; **14**: 198–208.
10. Thompson SG, Higgins JPT. Can meta-analysis help target interventions at individuals most likely to benefit? *Lancet* 2005; **365**: 341–346.
11. Federal Judicial Center. *Reference Manual on Scientific Evidence*, 2nd edn. Federal Judicial Center, Washington, DC, 2000, p539.
12. Reidenberg MM, Lowenthal DT. Adverse nondrug reactions. *N Engl J Med* 1968; **279**: 678–679.
13. Meyer FP, Troger U, Rohl FW. Adverse nondrug reactions: an update. *Clin Pharmacol Ther* 1996; **60**: 347–352.
14. Moore RA, McQuay HJ. Prevalence of opioid adverse events in chronic non-malignant pain: systematic review of randomized trials of oral opioids. *Arth Res Ther* 2005; 7: R1046–R1051.
15. Higgins PD, Johanson JF. Epidemiology of constipation in North America: a systematic review. *Am J Gastroenterol* 2004; **99**: 750–759.
16. Ziegler DK, Mosier MC, Buenaver M, Okuyemi K. How much information about adverse effects of medication do patients want from physicians? *Arch Intern Med* 2001; **161**: 706–713.
17. Yang YX, Lewis JD, Epstein S, Metz DC. Long-term proton pump inhibitor therapy and risk of hip fracture. *JAMA* 2006; **296**: 2947–2953.
18. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; **312**: 71–72.
19. Grimley Evans J. Evidence-based or evidence-biased medicine? *Age Ageing* 1995; **24**: 461–463.
20. Moore RA, Edwards JE, McQuay HJ. Acute pain: individual patient meta-analysis shows the impact of different ways of analysing and presenting results. *Pain* 2005; **116**: 322–331.
21. Kroenke K, West SL, Swindle R, *et al*. Similar effectiveness of paroxetine, fluoxetine and sertraline in primary care. *JAMA* 2001; **286**: 2947–2995.
22. Fries S, Grosser T, Price TS, *et al*. Marked interindividual variability in the response to selective inhibitors of cyclooxygenase-2. *Gastroenterology* 2006; **130**: 55–64.
23. Klepstad P, Dale O, Skorpen F, Borchgrevink PC, Kaasa S. Genetic variability and clinical efficacy of morphine. *Acta Anaesthesiol Scand* 2005; **49**: 902–908.
24. Lötsch J, Geisslinger G. Current evidence for a genetic modulation of the response to analgesics. Pain 2006; **121**: 1–5.
25. Dahlof CG, Pascual J, Dodick DW, Dowson AJ. Efficacy, speed of action and tolerability of almotriptan in the acute treatment of migraine: pooled individual patient data from four randomized, double-blind, placebo-controlled clinical trials. *Cephalalgia* 2006; **26**: 400–408.
26. Goadsby PJ, Massiou H, Pascual J, *et al*. Almotriptan and zolmitriptan in the acute treatment of migraine. *Acta Neurol Scand* 2007; **115**: 34–40.
27. Moore RA, Edwards JE, Hopwood J, Hicks D. Imiquimod for the treatment of genital warts: a quantitative systematic review. *BMC Infect Dis* 2001; **1**: 3.
28. Chengappa KN, Hennen J, Baldessarini RJ, *et al*. Recovery and functional outcomes following olanzapine treatment for bipolar I mania. *Bipolar Disord* 2005; 7: 68–76.
29. Moore A, Edwards J, Barden J, McQuay H. *Bandolier's Little Book of Pain*. Oxford University Press, Oxford, 2003.
30. Oldman AD, Smith LA, McQuay HJ, Moore RA. A systematic review of treatments for acute migraine. *Pain* 2002; **97**: 247–257.
31. Finnerup NB, Otto M, McQuay HJ, Jensen TS, Sindrup SH. Algorithm for neuropathic pain treatment: an evidence based proposal. *Pain* 2005; **118**: 289–305.