Knowing when evidence is trustworthy

R Andrew Moore,[1] Sheena Derry,[1] Scott A Strassels[2]

Evidence-based medicine is often seen as something dry, formal, and statistical, often used to justify a proscriptive approach to medicine. A more attractive approach is to use our understanding of those aspects of studies that can mislead us to identify the evidence we can trust. Evidence can, and probably should, be based on patient-centred outcomes of importance to clinical practice. The particular issues differ somewhat between clinical trials, observational studies, adverse events, diagnosis, and health economics. Here we explore some of important criteria relating to evidence from randomised trials, either alone or in meta-analyses.

Introduction

There are different ways of approaching and using evidence-based medicine (EBM). One is the dry formal approach, generally statistical, often used to justify a proscriptive approach to medicine. Another is to use analytical methods more freely to identify the evidence we can trust and to recognise that which is likely to be wrong. There are circumstances in which experience and common sense tell us that there may be problems with 'evidence', and that we should proceed with caution.

The issues are somewhat different for clinical trials, observational studies, adverse events, diagnosis or health economics.[1] Here we explore some issues of importance to interpreting evidence from randomised trials, either alone or in meta-analyses.

Is most published research false?

It has been said that only 1% of articles in scientific journals are scientifically sound. Research findings are more likely to be false when:[2]

- studies are small;
- the effect size is small;
- a greater number and a looser selection of relationships are tested;
- a greater flexibility in designs, definitions, outcomes and analytical modes are tested;
- greater financial and other interests and prejudices are involved (including research grants or the promise of future research grants);
- the topic is 'hot'.

There are many potential pitfalls to be aware of when assessing evidence. It is all too easy to be misled by an apparently perfect study that later is shown to be wrong, or by a meta-analysis with impeccable credentials that seems to be trying to pull the wool over our eyes.

It is common for early outstanding results to be followed by others that are less impressive. For example, The *New England Journal of Medicine*, *JAMA* and *Lancet* were searched for studies published between 1990 and 2003, each with more than 1000 citations.[3] Forty-five of 49 articles claimed an intervention to be effective based on sample sizes as low as 9 (nine) and as high as 87 000. Seven of the 45 were contradicted by later research, including one case series with nine patients, three cohort studies with 40 000–80 000 patients and three randomised trials with 200, 875 and 2002 patients, respectively. Only 3/43 (7%) randomised trials were contradicted, compared with 1/2 (50%) case series and 3/4 (75%) cohort studies.

The lesson is that, if we accept evidence of poor quality without validity or where there are few events or numbers of patients, we are likely to be misled. If we concentrate on evidence of high quality which is valid and with large numbers of patients, it will hardly ever happen.

Statistical significance and multiple statistical testing

It is an unspoken belief (supported by studies of publication bias) that reporting a result that is statistically significant helps to get a paper published. This leads to significance chasing, where data are analysed with the aim of finding any relationship showing significance at the 5% level. 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. You might want to ask yourself how happy you are with 1 in 20. Consider throwing two dice; double six occurs not infrequently, and that is a chance of 1 in 36. Recognising significance only when it is at the 1 in 100 level (1% or a p value of 0.01) often changes the perspective of the results.

But size alone is not enough. Statistical significance can mislead when we do not use statistics properly. Multiple subgroup analyses are common in published articles in our journals, usually without any adjustment for multiple testing. Of 131 randomised trials published in top journals in 6 months in 2004, there was an average of five subgroup analyses and 27 significance tests for efficacy and safety.[4]

This large population-based retrospective cohort study[4] underscored the problems that multiple statistical tests can pose by linked administrative databases covering 11 million adult residents of Ontario who were alive and had celebrated a birthday in the year 2000. All hospital admissions classified as urgent were examined to determine which of these were admitted within the 365 days following their birthday, the diagnosis on admission 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.

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, suggesting that astrological sign was a determinant of hospital admission. However, correcting for all 14 718 comparisons used meant using a significance level of 0.000003 rather than 0.05, and this produced the expected result of no association between astrological sign and hospital admission.

[1]Pain Research and Nuffield Division of Anaesthetics, University of Oxford, The Churchill Hospital, Oxford, UK
[2]Division of Health Outcomes and Pharmacy Practice, College of Pharmacy, University of Texas at Austin, Austin, Texas, USA

Correspondence to Dr R A Moore, Pain Research and Nuffield Division of Anaesthetics, Nuffield Department of Clinical Neurosciences, University of Oxford, The Churchill Hospital, Oxford OX3 7LE, UK;
andrew.moore@ndcn.ox.ac.uk

Received 24 September 2012
Accepted 28 September 2012

Importance of size

There may be times when *any* statistical testing is inappropriate. When can we be sure that we have enough information to be sure of the result, using the mathematical perspective of sure—namely, the probability that we are not being misled by the random play of chance? This is not a trivial question given that many results—especially those concerning rare but serious harm—are driven by very few actual events.

A group from McMaster University proposed that, with <200 outcome events, research is useful only for summarising information and generating hypotheses for future testing.[5] A different approach, using simulations of clinical trials and meta-analyses, arrived at much the same conclusion—that with <200 events, the magnitude and direction of an effect becomes increasingly uncertain.[6]

The number of events needed to be reasonably sure of a result when event rates are low (as in the case for rare but serious adverse events) has been tested mathematically.[7] Simulating clinical trials involved varying event rates in experimental and control groups, using different probability limits of 5% and 1% (p values of 0.05 and 0.01) and using larger and smaller studies. Lower event rates and smaller differences in event rates between groups combined with a greater need to detect a difference and using p values of 0.01 rather than 0.05 all pushed the requirements of studies towards needing to detect more events and study larger numbers of patients. Once event rates fall to about 1% or so and differences between experimental and control also fall to less than 1%, the number of events needed approached 100 and the number of patients rises to tens of thousands.

This points to the inescapable conclusion that, with few events, our confidence in any result is highly compromised. As a rule of thumb, we can probably dismiss studies with <20 events, should be very cautious of those with 20–50 events and can be reasonably confident of studies with >200 events—if everything else is in order. However, small trials are also problematical because of the potential lack of rigour leading to methodological biases. So meta-analyses of small studies (using numbers per treatment arm of <100 to denote small) give better results than larger trials,[8] probably because larger studies are better conducted.

Subgroup analyses

Most studies use 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 divide the data, the fewer the number of actual events in each portion. In addition, creating subgroups can remove the benefits of randomisation in clinical trials. Subgroups almost always introduce the danger of some unknown confounding.

A good example of the danger of subgroup analysis giving rise to unknown confounding is found in a review article examining the 30-day outcome of death or myocardial infarction from a meta-analysis of platelet glycoprotein inhibitors.[9] Subgroup analysis indicated a highly significant ($p < 0.0001$) benefit in men but not women. Actually, men had higher levels of troponins (a marker of myocardial damage) than women, and when this was taken into account the difference was understandable, with more effect where there was greater myocardial damage; sex was not the source of the difference.

Imputation methods

What happens when a patient withdraws from treatment? The statistical response is to 'impute' the result at the end of the study, usually using a last-observation-carried-forward (LOCF) approach. When the proportion of patients withdrawing during the study is small this approach has little effect on the estimate of efficacy but, in many circumstances such as mental health or chronic pain, withdrawal rates over 12 weeks can be 50% or even more. In that circumstance, LOCF imputation can give a false impression of a drug's efficacy, especially if many withdrawals are due to adverse events.[10] For some drugs and conditions—for example, opioids in chronic non-cancer pain—an alternative approach analysing 'true responders' (patients with good pain relief and able to tolerate adverse events) shows that the drugs are only judged effective because LOCF was used.[11 12]

Importance of the individual patient

It is widely understood that not every patient with a particular condition benefits from treatments that are known to work on average. A clinical trial may tell us that 50% of patients benefit with drug compared with 20% with placebo, and we applaud a good number needed to treat (NNT) of 3.3. Yet that obscures the fact that half the patients do not have benefit and may have adverse effects.

A trial in which patients with depression were randomised to receive one of three antidepressants demonstrates how different we all are. The three drugs were, on average, the same.[13] Patients initially randomised to one treatment frequently changed to another; by 9 months only 44% were still taking the treatment to which they had been randomised. Some (about 15%) were lost to follow-up after baseline or when on any of the randomised 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 from those initially randomised. Patients and their doctors found the balance that was right for them between benefit and tolerable adverse events; almost 70% had a good outcome over the 9 months of the trial.

The degree of variability between individuals in their physiological response to drugs is remarkable, and best exemplified by a study of 50 healthy young volunteers receiving rofecoxib, celecoxib or placebo who underwent a battery of clinical pharmacology tests.[14] There was great variability between individuals, ranging from 50 to several hundred-fold in activity in different in vitro tests. 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 metabolising morphine, opioid receptors and blood–brain barrier transport of morphine by drug receptors all contribute to considerable variability between individuals in response to morphine.[15] A number of mechanisms can influence individual responses to analgesics.[16]

The practical implications of these findings relate particularly to the potential harm of overly limited formularies. They also challenge us to consider how to make decisions about individual patients when all we have are average results and no patient is 'average'.

Outcomes

The outcomes chosen for measurement or for reporting in trials are frequently inadequate or unhelpful, with little clinical utility. Ideally, a satisfactory outcome should involve both a benefit judged to be valuable by patients and tolerability, because adverse events are often a cause of discontinuation of an otherwise effective therapy.

In migraine, for example, the outcome 'mild or no pain at 2 h after therapy' was standard. This changed to 'no pain at 2 h' and then to 'no pain at 2 h plus no recurrence or need to use analgesics over the next 24 h'. 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 h and had no adverse effects;[17] this amounted to no more than 22% of the total, only 12% more than with placebo.

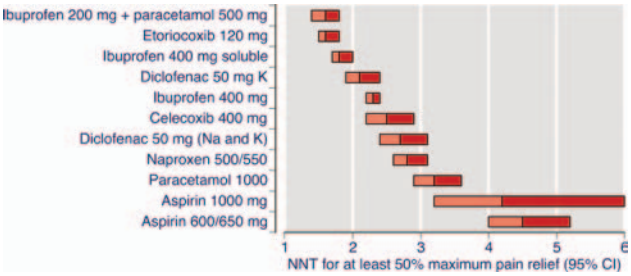

Figure 1 Efficacy of single-dose oral analgesics in acute pain following third molar extraction. Each drug/dose combination is from a Cochrane review, and the bars represent the 95% CI of the number needed to treat (NNT). The colour change shows the point estimate.

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,[18] while a longitudinal survey of patients with bipolar disorder suggested that, to be clinically relevant, success should be judged over longer periods because of the sustained nature of the disorder.[19]

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 encourage better use of what treatments we have and provide an incentive to develop better treatments for the future.

Conclusion

EBM is about a number of things, but first and foremost it is about avoiding being misled. To do this we need a passing acquaintance with issues of quality, validity and size, as well as how data are handled before being presented to us. When a trial uses tiny numbers and reports a subgroup analysis as statistically significant, especially using LOCF imputation, we should question the result and not rush to change practice.

EBM is also about improving outcomes for patients. This might involve defining better or more meaningful outcomes, 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.

Thirdly, when we collect together all the good evidence on a topic, we see the more clearly by eliminating misleading data. A number of examples exist in pain (especially in acute pain[20] and migraine[21]) and also in depression.[22] For example, figure 1 shows the NNTs for single-dose oral analgesics after third molar extraction from a Cochrane overview,[20] which has similar data on 46 drug/dose combinations for effective drugs in postoperative pain and much more data on drugs where there is no effect or where the evidence is inadequate. When we have good evidence—evidence that meets all the criteria of quality—then it becomes reliable and trustworthy, doing just what it says on the tin![1,23]

The final message should be about the importance of wisdom. In its fullest sense, EMB should incorporate evidence, from whatever source, with 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.

Competing interests None.

Provenance and peer review Commissioned; internally peer reviewed.

Key messages

- Much of the evidence we see is at best misleading and may be wrong.
- Understanding some relatively simple rules of evidence can help us to spot evidence that is not trustworthy.
- This might be studies that are small, multiple statistical testing or subgroup analyses without correcting for multiple testing, usingoutcome measures that are not patient or practice orientated, using inappropriate imputations when patients withdraw from treatment, or using average results when few patients are average.
- The good news is that there are examples of trustworthy evidence we can use, and their number is growing.

References

1. **Moore RA**, McQuay HJ. *Bandolier's little book of understanding the medical evidence*. Oxford: Oxford University Press, 2006.
2. **Ioannides JPA**. Why most published research findings are false. *PLoS Med* 2005;**2**:e124.
3. **Ioannides JPA**. Contradicted and initially stronger effects in highly cited clinical research. *JAMA* 2005;**294**:218–28.
4. **Austin PC**, Mamdani MM, Juurlink DN, *et al*. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol* 2006;**59**:964–9.
5. **Flather MD**, Farkouh ME, Pogue JM, *et al*. Strengths and limitations of meta-analysis: larger studies may be more reliable. *Control Clin Trials* 1997;**18**:568–79.
6. **Moore RA**, Gavaghan D, Tramer MR, *et al*. Size is everything—large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. *Pain* 1998;**78**:209–16.
7. **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.
8. **Nüesch E**, Trelle S, Reichenbach S, *et al*. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. *BMJ* 2010;**341**:c3515.
9. **Thompson SG**, Higgins JPT. Can meta-analysis help target interventions at individuals most likely to benefit? *Lancet* 2005;**365**:341–6.
10. **Moore RA**, Straube S, Eccleston C, *et al*. Estimate at your peril: imputation methods for patient withdrawal can bias efficacy outcomes in chronic pain trials using responder analyses. *Pain* 2012;**153**:265–8.
11. **Lange B**, Kuperwasser B, Okamoto A, *et al*. Efficacy and safety of tapentadol prolonged release for chronic osteoarthritis pain and low back pain. *Adv Ther* 2010;**27**:381–99.
12. **Steiner DJ**, Sitar S, Wen W, *et al*. Efficacy and safety of the seven-day buprenorphine transdermal system in opioid-naïve patients with moderate to severe chronic low back pain: an enriched, randomized, double-blind, placebo-controlled study. *J Pain Symptom Manage* 2011;**42**:903–17.
13. **Kroenke K**, West SL, Swindle R, *et al*. Similar effectiveness of paroxetine, fluoxetine and sertraline in primary care. *JAMA* 2001;**286**:2947–95.
14. **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.
15. **Klepstad P**, Dale O, Skorpen F, *et al*. Genetic variability and clinical efficacy of morphine. *Acta Anaesthesiol Scand* 2005;**49**:902–8.
16. **Lötsch J**, Geisslinger G. Current evidence for a genetic modulation of the response to analgesics. *Pain* 2006;**121**:1–5.
17. **Dahlof CG**, Pascual J, Dodick DW, *et al*. 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**:400–8.
18. **Moore RA**, Edwards JE, Hopwood J, *et al*. Imiquimod for the treatment of genital warts: a quantitative systematic review. BMC Infect Dis 2001;1:3.
19. **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.
20. **Moore RA**, Derry S, McQuay HJ, *et al*. Single dose oral analgesics for acute postoperative pain in adults. *Cochrane Database Syst Rev* 2011;**9**:CD008659.
21. **Oldman AD**, Smith LA, McQuay HJ, *et al*. A systematic review of treatments for acute migraine. *Pain* 2002;**97**:247–57.
22. **Cipriani A**, Furukawa TA, Salanti G, *et al*. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet* 2009;**373**:746–58.
23. **Moore RA**, Derry S. Why evidence matters. In: Stannard C, Kalso E, Ballantyne J, eds. Evidence-based chronic pain management. Oxford: Wiley Blackwell, 2010, 3–13.