Bandolier Extra

Evidence-based health care **February 2002**

EVIDENCE AND DIAGNOSTICS

Summary

This essay was developed from thoughts on the evidence-base of diagnostic testing arising from writing Bandolier. We can do great things in understanding treatments, both those done in the past, and planning trials to be done in the future. Did diagnostic testing match up?

- Most diagnostic tests are evaluated using architecture subject to immense bias
- Few systematic reviews of diagnostic tests are useful, because they just summarise wrong results
- For many tests there is too little information about how to use them, and when that has been examined it demonstrates massive lack of agreement
- In most circumstances we need to start afresh with new, better, and more directed research
- There are great paradigms for us, in the Ottawa ankle and knee rules, and the CARE study
- Diagnostic testing is a source of major economic waste in health services
- Healthcare systems avoid tackling the problems at their peril. Accurate and fast diagnosis is the key to accurate, fast, and cost-effective treatment. Major research investment cannot be avoided, but diagnostic research can cement relationships and be a driver for better use of knowledge and evidence

"Evidence-based medicine is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients." [1]

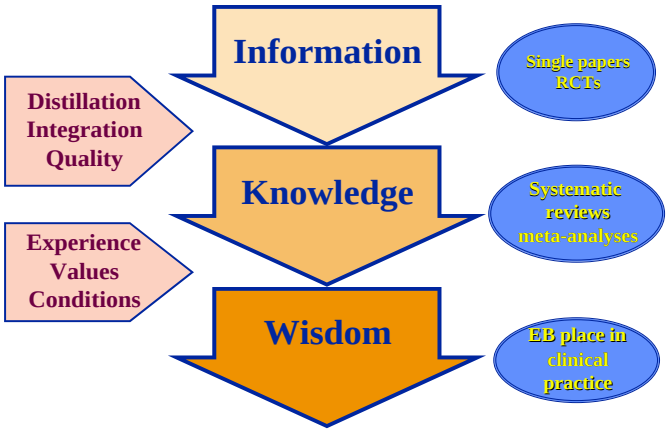
This quotation from Dave Sackett and his colleagues is as good a place as any to start thinking about evidence and migraine trials and treatments. The full article goes beyond this definition and includes patient and societal values. The main issue, though, is about where the practitioner goes to find "current best evidence". It could be using local or national guidelines, as for instance produced by organisations like the National Institute of Clinical Excellence, or those produced by eminent bodies. Some people will remain sceptical, though, and will (and should) satisfy themselves that the evidence on which guidelines are based is sound.

Information, knowledge and wisdom

In the past that task was difficult. With millions of papers being published each year (there is said to be about 30,000 medical journals), trying to find information, especially all the information was a heroic task. Now it is much easier. We can search PubMed online or visit electronic journals like BioMed, or electronic versions of paper journals like the BMJ. The Cochrane Library, available online or on CD for a small subscription has not only many good reviews, but also has over 250,000 controlled trials found by hand-searching the literature.

Good systematic reviews are increasingly available, where someone has asked a clinical question, and then summarised all the known information into a solid piece of knowledge. In doing so they will distil the information, perhaps integrate different types of information, and use quality filters so that only the most reliable information is used and that unreliable information is discarded.

How that knowledge is used depends on the practitioner making the conscientious, explicit and judicious use of the knowledge, in terms of the unique biology of the patient, incorporation patient concerns, their own experience and local knowledge, the values of society and the conditions in which they are working. The same piece of knowledge will play differently in Cardiff or Calcutta. That's the wisdom bit. That is why evidence-based approaches have nothing to do with rules, but should be seen as tools to allow practitioners to be better, and patients to be better informed.



Bias in clinical trials

One of the things we have learned through doing systematic reviews (also called meta-analysis when we pool data and do some sums) has been that certain types of study architectures are likely to produce results that are more favourable to a new treatment than they should be [2]. This is called bias, and many forms of bias have been discovered. We know that trials that are not randomised over-estimate the size of a treatment effect, as do trials that are not blind, or where information from patients is duplicated [3], or where trials are small [4], or where they have poor reporting quality [5,6].

Bias: a one-sided inclination Of the mind

	Over-estimation of treatment effect
Not random	40%
Not double-blind	17%
Duplicate information	20%
Small trials	30%
Poor reporting quality	25%

We can be much more specific. For instance, in a study of transcutaneous electrical nerve stimulation in postoperative pain, 17 of 19 trials that were not randomised came up with a positive result, while 15 of 17 randomised trials came up with the completely opposite result, that it did not work [7].

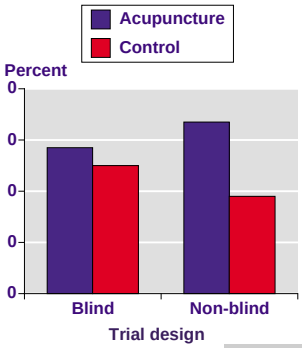
TENS in Postoperative Pain

	Analgesic result Positive	Negative
Randomised	2	15
Inadequate or not randomised	17	2

Carroll et al BJA 1996; 77: 798-803

In a review of acupuncture in back pain [8], lumping together all randomised trials, whether blinded or open, came up with the result that acupuncture worked for back pain. When you look at the open studies, where people making the assessments knew who had true acupuncture and who did not, there was a striking difference. But when you look at only the blinded studies, where people making the assessments did not know the treatment used, there was no difference at all. Acupuncture does not work.

Acupuncture in back pain



- Randomised studies
- Compared with sham acupuncture
- Validity of acupuncture checked independently
- 12 trials in total, 9 with outcomes
- 4 were blind (250 patients)
- 5 were not blind (204 patients)

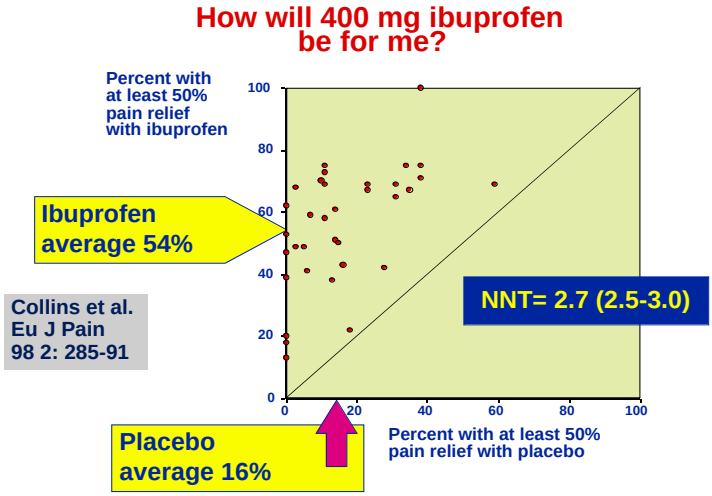
Ernst & White. Arch Intern Med 1998 158:2235-41

So attending to bias is an important issue in systematic reviews or meta-analysis of treatments. Where bias is known or likely to exist, then we may come up with the wrong overall result. To be sure of what we conclude in terms of best evidence, we have to use knowledge that is the very best. If we use poor quality knowledge, we may end up doing the wrong thing.

There are also some important issues around trial validity, summarised for acupuncture [9].

Size (bigness, magnitude)

We also have to be sure that we have sufficient information on which to base a conclusion. The figure below looks at all the literature available on properly randomised, double-blind trials comparing ibuprofen with placebo in acute postoperative pain [10]. They were impeccable trials, all using the same patients with the same initial degree of pain, and used the same outcomes over the same period of time.

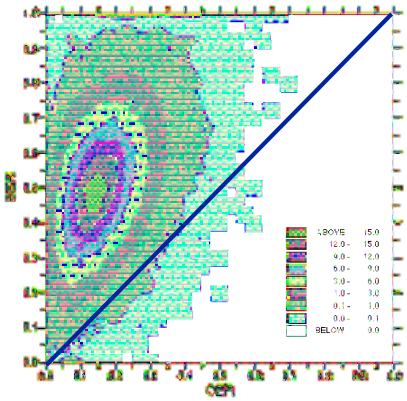


Each point represents a trial, and we plot the percentage with at least to% pain relief with placebo on the bottom, and the percentage with at least 50% placebo with ibuprofen 400 mg on the Y axis. All are above the line of equality, showing that ibuprofen is a better analgesic than placebo, which is encouraging. We can even see that the NNT of 2.7 means that ibuprofen is an effective analgesic.

But why do we have such a scatter of points if all these trials are supposed to be the same. Is it because some were conducted in Welsh wimps and others in Scottish stoics, perhaps? Actually, no. These trials were all done to show that ibuprofen is better than placebo. They had about 40 patients per treatment group to do this. They were not done to show how much better ibuprofen is than placebo, a subtly different question, and one that needs far more patients to answer accurately.

Because we know how over 5,000 individual patients perform in these trials, we can mathematically model the effects of the random play of chance on these trials. In the representation below [11], anywhere in the grey area is where a trial comparing ibuprofen 400 mg with placebo could fall just by chance. It is more likely to be in the redder areas, but the spread we see because of chance is at least as big as that we saw in practice with all the randomised comparisons of ibuprofen with placebo. So we don't need to seek abstruse reasons for differences between single trials until the effects of random chance have been eliminated. Only numbers will do that.

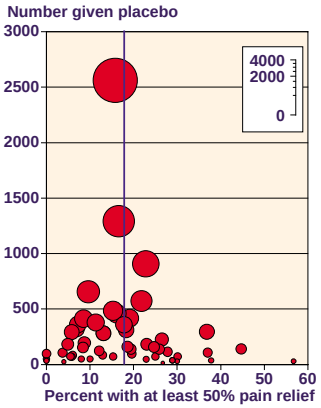
Simulated L'Abbé plot for ibuprofen-placebo comparison with mean of 40 patients per group



Moore et al. Pain 98 78: 209-16

Just to finish off the business of size, and to emphasise again how important it is, the slide below is probably unique in that it draws together information from of 50 meta-analyses [12]. Each blob represents the response rate found with placebo. We are plotting the rate or people achieving half pain relief with placebo against the number of patients given placebo. In total there are 12,000 such patients, and the blue vertical line represents the overall response rate of 18%. Only when the number of patients with placebo in the meta-analysis is large (of the order of 1000), is the overall rate accurately measured. This emphasises that size is everything.

Size really is everything



- Each symbol is percent with at least 50% pain relief with placebo in a systematic review
- Inclusion criteria and outcomes identical for top quality trials
- Below 500 patients average placebo response varies widely - between 0% and 56%
- Overall average is 18%

Moore. IASP 2000

Evidence and bias in diagnostic testing

For treatments, people have devised various levels of evidence, and this has been done in a number of other areas. The aim is to try to help us to use the best available evidence in making our decisions. One of the best places to see some thoughtful stuff is at the Centre for Evidence-Based Medicine. Usually at the top level is a systematic review of qualitatively good studies. But there are problems with this, because we may not always be able to recognise what constitutes goodness. Another set of levels of evidence in diagnostic testing uses criteria set out for individual studies of diagnostic tests (Table 1).

The top level is taken up by studies which have independent, blinded comparisons of the test with a reference standard, using consecutive patients. Other study architectures are given a lower level of evidence. Level 2 is the same as level 1, but using non-consecutive patients, for instance test-

Table 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 non-consecutive 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.

Table 2: Effect of different quality criteria on relative diagnostic odd ratios

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

ing the test on a group of people with the disease and a group of people without the disease, the most common study architecture. The problem for us is that we do not always recognise how big this difference is, and whether lower levels of evidence are so low as to mean that we can ignore them.

A review from Holland [13] gives us a real insight into the size of the gap between level 1 and level 2 studies. It 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 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 studies, only 15 of which satisfied all eight criteria of quality for the analysis. Thirty percent fulfilled at least six of eight criteria. The relative diagnostic odds ratio used indicated the diagnostic performance of a test in studies failing to satisfy the methodological criterion relative to its performance in studies with the corresponding feature. Over-estimation 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 (Table 2).

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.

The size of the bias is rather large, and tells us that if we use studies that compare people with the disease with those who do not have it, the results we get will be wrong. They will massively over-estimate the effectiveness of the test. That effectiveness will also be over-estimated by a range of other architectural problems.

Our problems with the quality of data from diagnostic test papers is compounded by how poorly they are reported. A study by Read and colleagues in 1995 [14] examined issues of quality of reporting of diagnostic tests (Bandolier 26). It described seven quality criteria, and then explored how those criteria were met in papers on diagnostic testing published by the four major English-language medical journals. The results were not encouraging: few told us anything useful about the patients being tested, and only a quarter told us how reliable and reproducible the test was (Table 3).

It is immediately clear, then, that for diagnostic testing the strategy of performing systematic reviews may just not be helpful. We would hesitate to base major decisions on trials of treatment that were known to have massively biased results, and yet for diagnostic testing that's usually all we have. For some major areas of medicine one can start with several thousand papers on diagnostic tests, and end up with fewer than a handful that might be included in a review. We really don't know very much that's any use about almost any test.

Systematic review should be about picking the nuggets of gold out of the dross. It is not about heaping small piles of dross into one big pile of dross.

Size and diagnostic tests

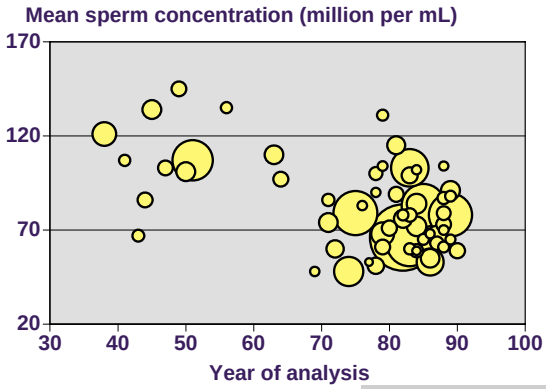
This is an issue that has probably not been addressed sufficiently. To explain how important size is, let's take the example of sperm counts [15,16]. Everyone knows that sperm counts are falling, and the reasons might include tight underpants, or oestrogens in the water supply, or even feminism. The evidence comes from a review of sperm counts (see Bandolier 56). This showed that sperm counts earlier in the century were higher than sperm counts later in the century.

The problem was that the early data came from a few studies with small numbers of men. If we re plot the data with the symbols properly related to the size of the individual study, we get a very different picture.

Table 3: Standards of reporting quality for studies of diagnostic tests

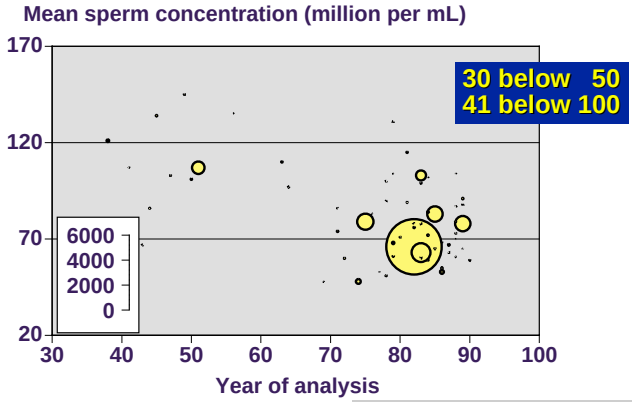
Reporting standard	Background	Criteria	Percent meeting standard
Spectrum composition	The sensitivity and specificity of a test depend on the characteristics of the population studied. Change the population and you change these indices. Since most diagnostic tests are evaluated on populations with more 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 sub groups. For successful use of the test, separate indices of accuracy are needed for pertinent individual sub groups within the spectrum of tested patients.	This standard is met when results for indices of accuracy were reported for any pertinent demographic or clinical sub group (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
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

Sperm counts are falling



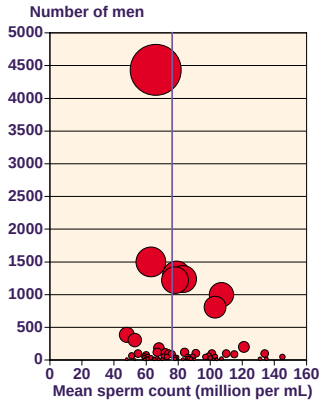
Carlsen et al. BMJ 92 305: 609-13

Or are they?



Carlsen et al. BMJ 92 305: 609-13

Size really is everything



- Six studies (>1000) with 10,696 men have weighted mean sperm count of 74 million/mL
- 14 studies (100-1000) with 2,844 men have weighted mean sperm count of 84 million/mL
- 41 studies (<100) with 1,407 men have weighted mean sperm count of 83 million/mL
- Overall average is 77

The simple fact is that the overall sperm count in the review, weighted by study size, was 77 million per mL. Only large studies with at least 1000 men came close to measuring it accurately, and small studies had values with averages from 30 to 140 million per mL.

And this is before we get to the point about how to measure sperm, what is a sperm, and what quality control between laboratories looks like. There is some suggestion in the literature that individual laboratories very widely in the results they give to the same sample.

The plain fact is that there is no evidence that sperm counts are falling. All the large (and good) studies give the same result. In the meantime, your tax is being used to finance research to find out if sperm counts are falling, how fast they are falling, and why they are falling. What a waste!

Good tests can make a difference

There are examples where a test and a treatment come together to make a difference. Examples include:

- Testing for Helicobacter pylori and effective treatments to eradicate ulcers.
- Using gene amplification methods for Chlamydia combined with azithromycin in screening.
- Measuring HIV viral load (and here) with protease inhibitors to make a real difference.

There are probably many more, but one problem that Bandolier has had is finding them. The evidence-base for effective diagnostics or diagnosis is rather thin - some would say pitifully thin. Think for a moment that effective treatment depends on effective diagnosis, and it makes one a bit concerned about the efficiency of our health services.

Not all tests are good, though

We must not delude ourselves that all tests are helpful. In pathology, the agreement between individual pathologists is not good, and even experts on the same disease can disagree frequently when looking at the same slides down a microscope (Bandolier 37; [17]). In reviewing 37 cases of possible melanoma (albeit not the easiest), eight benign cases and five malignant cases were agreed unanimously. Lack of unanimous agreement occurred in 24 cases (62%). Two or more discordant diagnoses were made in 14 cases (38%) and discordance was three or more in 8 cases (22%). The kappa was 0.5, indicating only moderate agreement.

It was illuminating to look at the extremes. One expert (and these were all experts in melanoma, don't forget) thought 21 cases were malignant and 16 were benign. Another thought 10 were malignant, 26 benign, and one indeterminate. Between them, these two pathologists disagreed on 12 out of 37 cases, and in 11 cases one pathologist identified a case as malignant while the other identified the same case as benign.

This is not picking on pathologists. We could make comments about other laboratory tests or imaging and its use in certain circumstances, like PSA for screening for prostate cancer (Bandolier 26), or imaging the back [18]. The point is that we need to know how good a particular test is for a particular patient at a particular level of suspicion. The book by Dave sackett and colleagues, "Evidence-based Medicine, how to practice and teach EBM", published by Churchill Livingstone, is a must for better understanding of testing.

How doctors use tests

Another cracking study [19] was dealt with in Bandolier 61. It asked groups of about 50 physicians and surgeons how they used diagnostic tests. The results were that very few knew or used Baysean methods, or ROC curves, or likelihood ratios. So the formal ways we have of explaining diagnostic test results, including sensitivity and specificity, and just not understood or used by the people who use the tests.

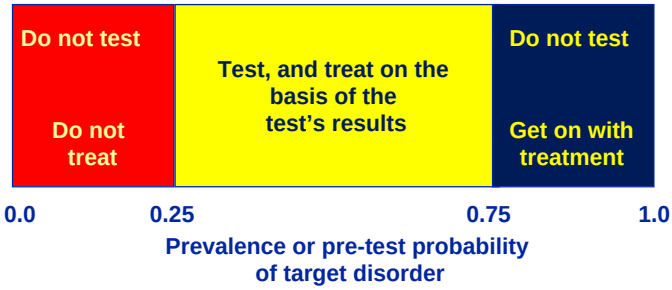
How do doctors use tests?

Specialisation	Bayesian method	ROC curve	Likelihood ratios
Specialist physician	5	1	1
Generalist physician	2	0	1
Paediatrician	1	1	0
General surgeon	0	1	0
Family practice	0	0	0
Obstetrics/Gynaecology	0	0	0
Overall percentage	3%	1%	1%

Frequency of use of methods of assessing test accuracy: 50 physicians in each category

Reid et al. Am J Med 98 104: 374-80

If asked, what most doctors want is not likelihood ratios or sensitivity, specificity or positive predictive value. They want simple algorithms, ideally on their PC or palm pilot, that can be used to help make decisions. Even simple ways of looking at likelihood ratios assumes you know where you are starting from.

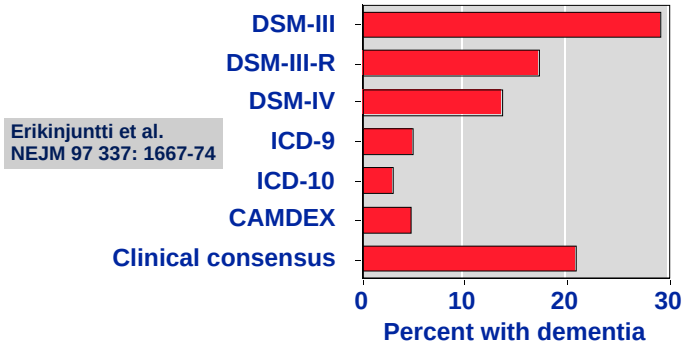


Clinical scoring systems are mixed blessings. A number of different scoring systems for dementia were examined by a multi-disciplinary team using detailed notes on just under 1900 patients [20]. The rate of dementia varied from 3% using one system to 30% using another.

If we can't diagnose dementia accurately, how can we do clinical trials, how can we measure success, or decide which patients benefit, or advise clinicians, or explain all this to patients or their families?

For thyroid function tests, a small survey in The Lancet almost 25 years ago [21] indicated that clinical scoring systems good give excellent predictors of when laboratory tests need to be done to confirm a diagnosis. If there are fewer than three signs or symptoms, the chances of thyroid disease are lower than the population in general, and knowing this could prevent 90% of all tests being sent to labs from GPs or outpatients.

Prevalence of dementia according to system of diagnosis in 1879 subjects



Erikinjuntti et al. NEJM 97 337: 1667-74

Symptom scoring can help

Signs and symptoms in primary care were thyroid (18), cardiac (6) and others (3)

Number of symptoms	Number	With thyroid disease	Percent
Five or more	23	18	78.00
Three or four	35	1	2.90
Two or fewer	442	2	0.45
Total	500	21	4.20

White & Walmsley, Lancet 1978 ii: 933-5

Doing better - CARE

So can we do better in how we think about and evaluate diagnostic testing? Sure we can. First of all we have some excellent examples of how we might go about evaluating tests. The best examples are the Ottawa ankle [22] and knee rules [23]. In each case what we had studies that:

- Collected good quality information on a learning data set to establish what clinical signs and symptoms were positively correlated with X-ray confirmed fracture
- Formulated a clinical decision rule based on the positively correlated factors
- Validated the rule on an independent data set to show that it worked
- Tested the rule in a randomised controlled trial to establish that it could be used, it would be used, and that using it made an important difference.

The result was a clinical decision rule that worked, and was used, provided a better service to patients, and saved time and money. Similar approaches have been made for rules for discontinuing cardiac resuscitation in hospital [24].

One of the most exciting new developments in e-medical research is that on the Clinical Assessment of the Reliability of the Examination (CARE), which is a collaborative study of the accuracy and precision of the clinical examination [25]. If you want to know all about it, it's Internet address is http://www.carestudy.com/.

Essentially a group of people get together via the Internet to contribute patients to studies of clinical diagnostic testing. A systematic review is first undertaken, and only those features most likely to be important are combined in the final protocol. Because doctors around the world were involved, they were able to collect information in 10 times more people than in the medical literature in just a few weeks, and come up with some simple decision rules that doctors find useful, for instance for chronic obstructive airways disease [26].

Their website is a must. It is simply one of the best things in the world for diagnostic testing. The scandal is that demand for better information is so great, and supply so limited. It's important, too, because testing consumes resources, and getting it wrong can be a disaster for patients and providers.

GPs order blood tests on one in every 25 patients they see. In hospital it's probably more. We know from stories carried in Bandolier that unnecessary tests [27] can be a large proportion of the total, with huge financial implications. We know that if these tests were not necessary according to guidelines, and potential savings in time and cost are possible [28,29]. We also know from a randomised trial that by having guidelines on GPs computers, we can actually reduce the number of tests ordered substantially [30].

The lesson is that doing simple things well makes for a better quality service at a lower cost, if for no other reason that doing fewer tests means fewer false positive results.

Where do we go from here?

If you are in a hole, stop digging. What we are doing now is so awful that we have to scrap most of it and start afresh. Doing systematic reviews of diagnostic tests is a complete waste of time.

- If we do new studies, they must be free from bias. Consecutive patients in real situations is the only way to test tests.
- Studies have to be large to avoid random error.
- Use the Internet, like CARE to recruit many centres and do things faster.
- Combine clinical AND laboratory findings.
- Choose clinical situations where doctors need most help.
- Find better ways of giving results - not normal ranges, but algorithms. Start work on new ways of expressing outcomes of diagnostic tests.
- Make results available through better electronic communication.

One thing is certain. This should be one of the most fertile areas for research in the next few years. Laboratory scientists, clinicians, nurses, pharmacists and others all should be able to take part. It doesn't all need a brain the size of a planet. It doesn't have to be done at some ivory tower, and much could be done in Grimsby on a wet Tuesday afternoon. Watch this space.

©Bandolier: 08-Jan-2002

References

1 DL Sackett et al. Evidence-based medicine: what it is and what it isnÕ British Medical Journal. 1996; 312:71-72.

2 KF Schulz et al. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. Journal of the American Medical Association. 1995; 273:408-412.

3 MR Tramr et al. Impact of covert duplicate publication on meta-analysis: a case study. British Medical Journal. 1997; 315:635-639.

4 RA Moore et al. Quantitive systematic review of topically-applied non-steroidal anti-inflammatory drugs. British Medical Journal. 1998; 316:333-338.

5 KS Khan et al. The importance of quality of primary studies in producing unbiased systematic reviews. Arch Intern Med. 1996; 156:661-666.

6 D Moher et al Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials. 1995 Feb; 16:62-73.

7 D Carroll et al. Randomization is important in studies with pain outcomes: Systematic review of transcutaneous electrical nerve stimulation in acute postoperative pain. British Journal of Anaesthesia. 1996; 77(6):798-803.

8 E Ernst & AR White. Acupuncture for back pain. Arch Intern Med. 1998; 158:2235-2241.

9 LA Smith et al. Teasing apart quality and validity in systematic reviews: an example from acupuncture trials in chronic neck and back pain. Pain. 2000; 86:119-132.

10 SL Collins et al. Oral ibuprofen and diclofenac in postoperative pain: a quantitative systematic review. European Journal of Pain. 1998; 2:285-291.

11 RA Moore 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; 78:209-216.

12 RA Moore. Understanding clinical trials: what have we learned from systematic reviews? Proceedings of the 9th World Pain Congress. 2000.

13 JG Lijmer et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999 282: 1061-6.

14 MC Read et al. Use of methodological standards in diagnostic test research: getting better but still not good. Journal of the American Medical Association 1995 274:645-51.

15 S Becker & K Birhane. A meta-analysis of 61 sperm count studies revisited. Fertility and Sterility 1997 67: 1103-8.

16 E Carlsen et al. Evidence for decreasing quality of semen during past 50 years. British Medical Journal 1992 305: 609-13.

17 ER Farmer et al. Discordance in the histopathologic diagnosis of melanoma and melanocytic nevi between expert pathologists. Human Pathology 1996 27: 528-31.

18 MC Jensen et al. Magnetic resonance imaging of the lumbar spine in people without back pain. New England Journal of Medicine 1994 331: 69-73.

19 MC Reid et al. Academic calculations versus clinical judgements: practicing physiansÕ use of quantitative measures of test accuracy. American Journal of Medicine 1998 104: 374-80.

20 T Erikinjuntti et al. The effect of different diagnostic criteria on the prevalence of dementia. New England Journal of Medicine 1997 337: 1667-74.

21 GH White & RN Walmsley. can the initial clinical assessment of thyroid function be improved? Lancet 1978 ii: 933-5.

22 I Stiell et al. Multicentre trial to introduce the Ottawa ankle rules for use of radiography in acute ankle injuries. British Medical Journal 1995 311: 594-7.

23 IG Stiell et al. Implementation of the Ottawa knee rule for the use of radiography in acute knee injuries. JAMA 1997 278: 2075-9.

24 C van Walgrave et al. Validation of a clinical decision aid to discontinue in-hospital cardiac arrest resuscitation. JAMA 2001 285: 1602-1606.

25 FA McAlister et al. Why we need large, simple studies of the clinical examination: the problem and a proposed solution. Lancet 1999 354: 1721-24.

26 SE Straus et al The accuracy of patient history, wheezing, and laryngeal measurements in diagnosing obstructive airway disease. JAMA 2000 283: 1853-1857.

27 C van Walraven & CD Naylor. Do we know what inappropriate laboratory utilization is? A systematic review of laboratory clinical audits. JAMA 1998 280: 550-8.

28 DH Solomon et al. Techniques to improve physiciansÕ use of diagnostic tests. JAMA 1998 280: 2020-7.

29 C van Walraven et al. Effect of population-based interventions on laboratory utilization. JAMA 1998 280: 2028-33.

30 MA van Wijk et al. Assessment of decision support for blood test ordering in primary care. A randomized trial. Annals of Internal Medicine 2001 134: 274-281.