

MINDSTRETCHER

HOW GOOD IS THAT TEST?

More than ever the practice of medicine and delivery of healthcare depends upon making an accurate diagnosis based on the use of diagnostic tests. These may be radiological (including ultrasonics or magnetic resonance imaging), laboratory based (including biochemistry haematology, bacteriology, virology immunology or genetics) or physiological (thermometer, dipstick, exercise stress tests).

How do you know how good a test is in giving you the answer you seek? What are the rules of evidence against which new (or existing) tests should be judged? We can have rules of evidence for treatments (*Bandolier* 12), so rules of evidence for tests shouldn't be too much to ask for.

Methodological standards

Bandolier has found a terrific paper [1] which sets out seven methodological standards for diagnostic tests. It then looks at papers published in *Lancet*, *British Medical Journal*, *New England Journal of Medicine* and *Journal of the American Medical Association* from 1978 through 1993 to see how many reports of diagnostic tests meet these standards (and for those who can't stand the suspense, the answer is not many!).

This paper is neither an easy nor a comfortable read. For those of us engaged in providing diagnostic tests this is a rude reminder of how little time we spend thinking and how much time "getting and spending". For those of us who use diagnostic tests, it is also a rude reminder of how much faith we often put into a number or opinion, perhaps without thinking of the weight that should be placed on that number

For these reasons, and many others, this is a paper to get from the library and read in full, and then to keep handy for later use. *Bandolier* summarises the paper here to make sure that we understand and appreciate it.

Seven standards

Standard 1: Spectrum composition

The sensitivity or specificity of a test depends on the characteristics of the population studied (see *Bandolier* 3). Change the population and you change these indices. Since most diagnostic tests are evaluated on populations with significant disease, the reported values for sensitivity and specificity may not be applicable to other populations, in which the test is to be used.

For this standard to be met the report had to contain information on three of these four criteria: age distribution, sex distribution, summary of presenting clinical symptoms and/or disease stage, and eligibility criteria for study subjects.

Standard 2: Pertinent subgroups

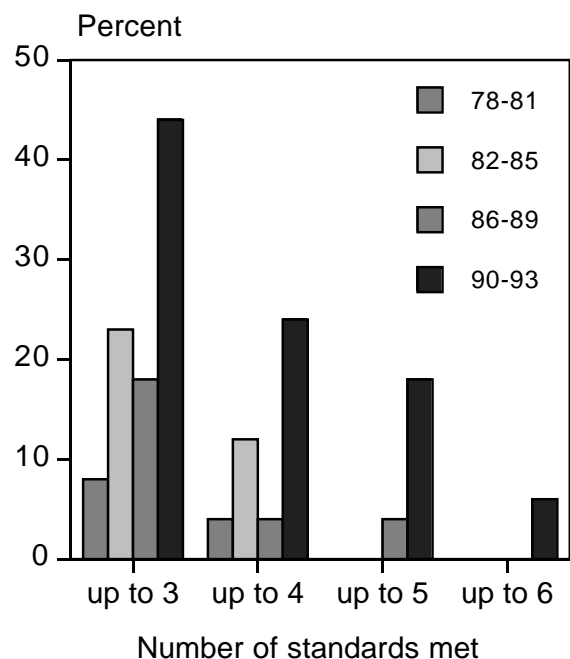
Sensitivity and specificity may represent average values for a population. Unless the condition for which a test is to be used is narrowly defined, then the indices may vary in different medical sub groups. For successful use of the test, separate indices of accuracy are needed for pertinent individual 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).

Standard 3: 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.

Proportion of reports of diagnostic tests meeting accepted methodological standards



The authors of the paper discuss this at length because of an early lack of agreement in applying the criteria for it. They give many examples. One was of a new DNA diagnostic test to detect the breast cancer gene administered to biopsy-positive breast cancer and to cancer-free controls. Since biopsy may be ordered preferentially in women with a family history of breast cancer, the group of "cases" will be enriched by a clinical factor which itself may be associated with the new DNA test.

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.

Standard 4: 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 stud-

ies or case-control studies, a statement was required regarding the independent evaluation of the two tests.

Standard 5: Precision of results for test accuracy

The stability 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.

Standard 6: Presentation of indeterminate test results

Not all tests come out with a yes or 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.

Standard 7: Test reproducibility

Tests may not always give the same result - for a whole vari-

Standard	Studies meeting the standard (%)
1: Spectrum composition (overall)	27
Age distribution	(55)
Sex distribution	(48)
Symptoms and/or stage	(29)
Study eligibility criteria	(20)
2: Accuracy in sub groups	9
3: Avoidance of workup bias	51
4: Avoidance of review bias	43
5: Test accuracy precision	12
6: Indeterminate test results	26
7: Test reproducibility	26

ety 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.

Do reports meet the standards?

Between 1978 and 1993 the authors found 112 articles, predominantly in radiological tests and immunoassays. Few of the standards were met consistently - ranging from 46% avoiding workup bias down to 9% reporting accuracy in subgroups (Table).

While there was an overall improvement over time for reports to score well on more standards, even in the most recent period studied only 24% met up to four standards, and only 6% up to six (Figure).

Implications

The authors suggest that, given that these reports were published in arguably the four most important medical journals in the world, they overestimate the true use of methodological standards in the evaluation and reporting of diagnostic tests. That may be, but even so, the findings give real cause for concern about the technological creep of diagnostic tests of unproven worth.

Systematic evaluation of diagnostic tests before their widespread use could be expected to provide benefits in several areas:-

- 1 Elimination of poor or useless tests before they become widely available.
- 2 Improved quality of diagnostic test information.
- 3 Reduced health care costs.
- 4 Improved patient care.

Given the reliance put on diagnostic tests in modern medical practice, it might be appropriate for diagnostic tests to be subjected to standardised evaluation before being released for widespread use [2]. This might seem draconian, but is there another way?

References:

- 1 MC Read, MS Lachs, AR Feinstein. Use of methodological standards in diagnostic test research: getting better but still not good. *Journal of the American Medical Association* 1995 274:645-51.
- 2 GH Guyatt, PX Tugwell, DH Feeny, RB Haynes, M Drummond. A framework for clinical evaluation of diagnostic technologies. *Canadian Medical Association Journal* 1986 134:587-94.

FUNGAL NAIL INFECTIONS

A GP called the *Bandolier* office and asked whether it was right that oral terbinafine was better than oral griseofulvin for fungal nail disease - onychomycosis. On occasions such as this the immediate response is "we don't know" - but followed by "we'll try and find out". One thing we found out early on was that this is a common complaint, affecting just under 3% of the population.

To try and find recent papers on this subject we searched MEDLINE using onychomycosis as a MeSH term, with terbinafine and griseofulvin as free text terms for the period June 1993 to December 1995. We found three randomised controlled trials and some cost-effectiveness analyses.

Randomised trials

One RCT examined short duration treatment for fingernail fungal infections and two looked at toenail infections. All were direct comparisons of terbinafine and griseofulvin. The three RCTs were of good quality. All established the fungal nature of the disease before the study began; although as many as 65% of patients were re-treatments, oral treatments had been discontinued for three months and topical treatments for one month before the study began.

Because of the slow-growing nature of nails, especially toenails, these studies take a long time: judgement of cure was usually not made until at least one year after the start of the treatment. All the studies had substantive numbers of patients in each treatment group, and two had power calculations to establish the size of the trial. All trials had excellent descriptions of withdrawals and adverse effects, and included patients who stopped therapy because of adverse effects or ineffective treatment in their analyses. For all three, a complete cure was defined as when there was no sign of clinically abnormal nail and when mycological cultures were negative.

Haneke et al [1]

This multicentre study from Germany was supported by Sandoz. Patients with fungal fingernail infections were randomly assigned to 250 mg/day oral terbinafine or 500 mg/day oral micronized griseofulvin. There was a 12 week active drug period, followed by a 12 week placebo period. Patients then entered a six month double-blind follow up phase. Outcome was judged at 48 weeks after the start of treatment - all of which had been conducted blind to which treatment was used.

Sixty-seven evaluable patients received terbinafine and 72 received griseofulvin. The clinical cure rates (mycological cure and free from clinical signs) were 51/67 and 28/72 respectively [odds ratio 4.5 (95%CI 2.3-8.8) and number-needed-to-treat (NNT) of 2.7 (95%CI 1.9-4.5)].

Hofmann et al [2]

This was also a multicentre study from Germany supported by Sandoz. Patients with fungal toenail infections (predominantly severe) were randomised to a 24-week treatment of

oral terbinafine 250 mg/day (followed by 24 weeks of placebo) or 48-week treatment of micronized griseofulvin 1000 mg/day

The chosen end points were 48 weeks and 72 weeks; the longer time point was to consider the effect of relapses. Eighty-three evaluable patients received terbinafine and 88 received griseofulvin.

At 48 weeks the clinical cure rates (mycological cure and free from clinical signs) were 56/83 and 49/88 respectively [odds ratio 1.6 (95%CI 0.9-3.0)]; there was no significant difference.

At 72 weeks the clinical cure rates (mycological cure and free from clinical signs) were 50/83 and 34/88 respectively [odds ratio 2.4 (95%CI 1.3-4.3) and NNT of 4.6 (95%CI 2.8-14)]. More relapses had occurred in patients treated with griseofulvin.

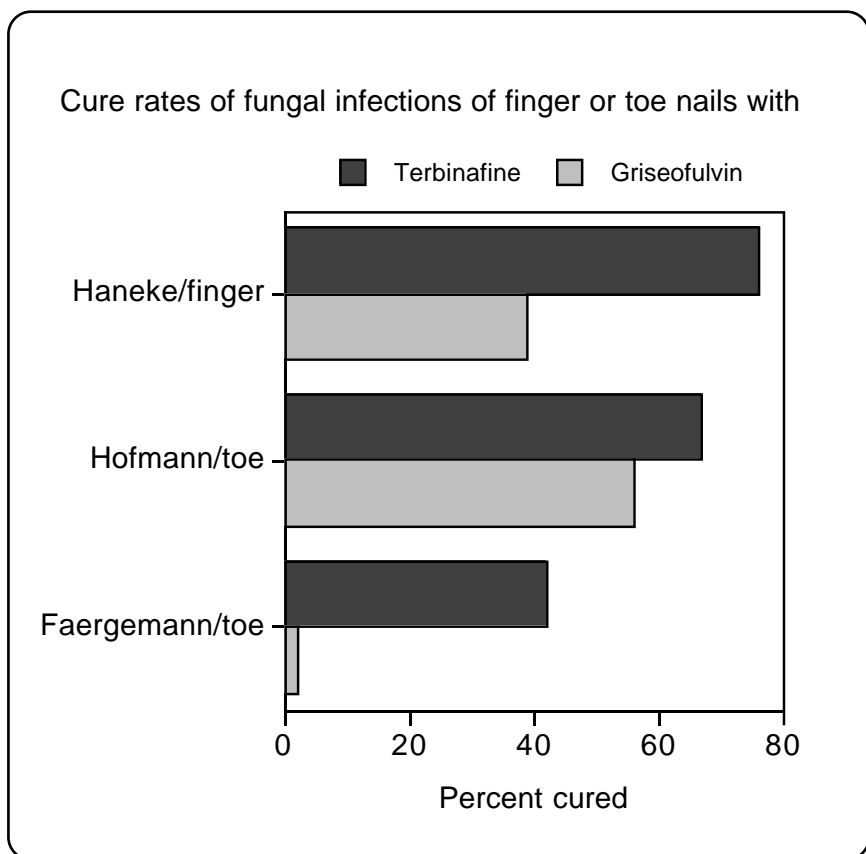
Faergemann et al [3]

This multicentre study from Sweden makes no comment about industrial support. Patients with fungal toenail infections (predominantly severe) were randomised to a 16-week treatment of oral terbinafine 250 mg/day (followed by 36 weeks of placebo) or 52-week treatment of micronized griseofulvin 500 mg/day

Forty-three evaluable patients received terbinafine and 41 received griseofulvin. The clinical cure rates (mycological cure and free from clinical signs) at 52 weeks were 18/43 and 1/41 respectively [odds ratio 9.2 (95%CI 3.4-25) and NNT of 2.5 (95%CI 1.8-4.2)].

Overall result

It is difficult and perhaps dangerous to combine data from different trials which use different doses, times and time of outcome. A clear picture does, however, emerge, where 119/193 patients were cured with terbinafine compared with 63/201 with griseofulvin, using a severe definition of a cure.



This would give an NNT of about three; that is, for three patients treated with terbinafine instead of griseofulvin, one *more* would be cured.

Adverse events

All three studies recorded adverse events scrupulously with possible drug-related adverse events being reported by about 40% of patients receiving either drug [1]. Adverse events were predominantly gastrointestinal, with headache also common for griseofulvin.

The numbers of drug-related study withdrawals in the three studies was 13/193 patients treated with terbinafine and 20/201 with griseofulvin.

Dose of griseofulvin

At least one of the studies [1] has been criticised [4] because the dose of griseofulvin recommended in the USA is 1000 mg/day; two of the studies used 500 mg/day. The dose was chosen because the German regulatory authorities would not allow a higher dose to be used. The Swedish study also makes the point that 500 mg/day of griseofulvin is the standard adult dose there.

Implications for costs

The cost of a daily dose of terbinafine is about 10 times greater than that of griseofulvin. This makes the drug cost of a course of treatment likely to be lower with griseofulvin, even though it has to be taken for much longer. Two pharmacoeconomic studies in a Canadian [5] or European [6] setting take into account the lower success rate and higher relapse rates seen with griseofulvin, and a different picture emerges.

The cost of treating a fungal infection of toenail or fingernail with griseofulvin is about 1.2 times more expensive than using terbinafine. The reason for this was that terbinafine was more effective in curing fungal nail disease, and there were fewer relapses after treatment had stopped.

These conclusions are also supported by a Dutch analysis - but as it was in Dutch we could only read the English abstract.

Comment

Study	Dose and time	Cured with Terbinafine	Cured with Griseofulvin	NNT
Fingernail [1]	T 250 mg/day 12 weeks G 500 mg/day 12 weeks	51/67	28/72	2.7 (1.9 - 4.5)
Toenail [2]	T 250 mg/day 24 weeks G 1000 mg/day 48 weeks	50/83	34/88	4.6 (2.8 - 14)
Toenail [3]	T 250 mg/day 16 weeks G 500 mg/day 52 weeks	18/43	1/41	2.5 (1.8 - 4.2)

This is obviously an area of great interest to GPs where practice may be changing quickly. There are earlier papers (with small numbers) looking at the terbinafine/griseofulvin comparison, and a simple meta-analysis of those was done for one of the pharmacoeconomic evaluations [6].

A short review of drug treatment for fungal nail disease [7] would be a useful read for most *Bandolier* readers as it simply puts over the background, alternative therapies, and their benefits and potential harm.

Bandolier would be pleased to hear from any Health Authorities or practices which have implemented evidence-based programmes for treatment of fungal nail disease in conjunction with dermatologists.

References:

- 1 E Haneke, I Tausch, M Bräutigam et al. Short-duration treatment of fingernail dermatophytosis: a randomized, double-blind study with terbinafine and griseofulvin. *Journal of the American Academy of Dermatology* 1995 32:72-7.
- 2 H Hofmann, M Bräutigam, G Weidinger et al. Treatment of toenail onychomycosis: a randomized double-blind study with terbinafine and griseofulvin. *Archives of Dermatology* 1995 131:919-22.
- 3 J Faergemann, C Anderson, Kjell Hersle et al. Double-blind, parallel-group comparison of terbinafine and griseofulvin in the treatment of toenail onychomycosis. *Journal of the American Academy of Dermatology* 1995 32:750-3.
- 4 DJ Hogan. Short-duration treatment of fingernail dermatophytosis: a randomized, double-blind study with terbinafine and griseofulvin. (letter and reply) *Journal of the American Academy of Dermatology* 1995 32:541-2.
- 5 TR Einarson, SR Arikian, NH Shear Cost-effectiveness analysis for onychomycosis therapy in Canada from a government perspective. *British Journal of Dermatology* 1994 130 (Supp 43): 32-4.
- 6 SR Arikian, TR Einarson, G Kobelt-Nguyen et al. A multinational pharmacoeconomic analysis of oral therapies for onychomycosis. *British Journal of Dermatology* 1994 130 (Supp 43): 35-44.
- 7 DT Roberts. Oral therapeutic agents in fungal nail disease. *Journal of the American Academy of Dermatology* 1994 31: S78-S81.

ASKING THE RIGHT QUESTIONS

A recent thought-provoking and useful article - "Never mind solutions: what are the issues?" [1] - raises many points relevant to *Bandolier* readers. The paper provides an insight into practical ways to link research and practice. It includes work undertaken in North Thames and draws on lessons from outside the NHS.

To date much of the evidence-based medicine lobby has operated through the notion of "technology push" *Bandolier's* own imagery of 'bullets' of evidence is consistent with this model. Dawson [1] reminds us of the lessons that innovation in practice can only happen when 'technology push' is coupled with 'user-pull'. This is not a new idea, though it is often overlooked. It is consistent with the views of Stocking in 1985 [2] when she wrote about the "importance of research findings being perceived as relevant and applicable by clinicians if they are to have a hope of being translated into practice".

The development of Clinical Effectiveness Units (for instance at Kettering General Hospital and Derbyshire Royal Infirmary) are designed to address this gap in perception by engaging clinicians in a process of continuous feedback linking R&D and audit. In any healthcare setting clinicians need to be involved in defining and re-defining the problems as well as implementing the solutions. This is a key strength of locally produced, evidence-based guidelines.

If, as Oxman concluded [3], there are no "magic bullets" for improving the quality of care, then we must be very sure that clinicians are included as an integral part of the process of identifying the targets or we may find ourselves aiming at the wrong target! If, in addition, we want to avoid shooting ourselves in the foot, we must consider all clinically relevant issues (including the social and organisational contexts of practice).

Rules for defining the problem

Clinicians in primary and secondary care swiftly challenge researchers, purchasers and professional advisers (eg Health Authority medical and pharmaceutical advisers, and District Drug and Therapeutics Committees) who fail to ask the right questions in pursuit of 'better patient care. So before of offering an evidence-based solution, we must always follow the rules for defining the problem:

- 1 Is it clear what issue is being addressed?
(why have we chosen this target?)
- 2 Is that issue clinically relevant?
(is it the right target?)
- 3 Have all the clinically relevant options been considered?
(are there other means of achieving similar or greater benefits - other targets - and have these other means been properly explored in the problem definition?)

Sue Ashwell
Director of Pharmacy Services, Kettering General Hospital

References:

- 1 S Dawson. Never mind solutions: what are the issues? Lessons of industrial technology transfer for quality in health care. *Quality in Health Care* 1995 4:197-203.
- 2 B Stocking. Initiatives and inertia: case studies in the NHS. London: Nuffield Provincial Hospitals Trust, 1985.
- 3 AD Oxman. No magic bullets. A systematic review of interventions to improve the performance of health care professionals. London: North East Thames RHA, R&D Directorate, 1994.

SCREENWATCH

Bandolier has received a number of letters saying how helpful readers have found our Screening Whitelist (approved screening programmes) and Screening Blacklist (programmes and tests which should not be used). We have also been taken to task for using "blacklist" as a pejorative term. Perhaps we were over-simple, for as a radiologist once said of mammographic screening, "it is not all black and white in radiology".

Scottish law is better

In English law one is innocent or guilty Scottish law has a third category not proven, which is as useful to epidemiologists as it is to lawyers. The three verdicts for screening epidemiologists are:-

- innocent** - doing more good than harm
- guilty** - doing more harm than good
- not proven** - of unknown effectiveness

Only the innocent should go freely into the health service, providing, of course, that they are also affordable at reasonable cost.

We have hitherto divided screening tests into black and white, guilty and innocent. The new Executive Letter on Clinical Effectiveness, signed by Alan Langlands, emphasises in an appendix that a number of interventions are currently the subject of research and should not be used except if they are part of a well-designed research project, and a number of these are screening tests.

Not proven, not innocent

Interventions are sometimes introduced into the service under the guise of "research" and then slip into routine clinical practice. This is certainly the case with some of the new genetic tests. The category of "not proven" allows us clearly to mark interventions which should be introduced only as part of a research study which has been through an Ethics Committee and is fully costed and supported. Most of these tests were on our Screening Blacklist, and in a simple binary view of the world should remain there and not be introduced as services. Only the innocent, interventions which have been shown to do more good than harm, should be introduced to the service.

PSA and screening

No Cancer **4.0** **Cancer**

< 4 µg/L

> 4 µg/L

95% CI

Interlaboratory variation

40% of men with prostate cancer are here

Ejaculation can result in a drop of 85%
in serum PSA levels
Simak et al, J Urol 93 306:895-7

PSA and prostate cancer screening

There may well eventually be useful screening programmes for prostate cancer, combined with effective remedies. The situation now is far from that. Just looking at a few of the issues, as in the figure, gives a taste of the problems.

For instance, even though the figure of 4 µg/L for PSA is commonly taken as an action limit, the 95% confidence interval for that measure in a typical laboratory is about 3.2-4.8 µg/L. If a sample with a "true" level of 4 µg/L were tested in every laboratory in the UK, the range might be as wide as 2.5-5.5 µg/L. This is not to denigrate the laboratories, since this would be typical within assay variation and between laboratory performance for this sort of test, with many different reagent kits and methods available.

It is also interesting to see what a literature search produces. Though there are perhaps over 1,000 papers published which include PSA measurements in the last three years, only two (so far as *Bandolier* could find) specifically addressed the effect of ejaculation.

One [1] involved men under 40 years (so may be of questionable value in older men) and showed that PSA values fell from a pre-ejaculation average of 1.4 µg/L to a one-day post-ejaculation value of 0.2 µg/L.

Another paper [2] found no significant relation between PSA levels and ejaculation; it looked at men in their '20s whose

median PSA was 0.6 µg/L (and mode 0.4 µg/L). It used a test which was not designed to measure PSA accurately below about 1 µg/L. So this study almost certainly lacked sensitivity to see any change, yet this reference has been taken by some urologists as a positive indication that ejaculation had *no* effect on serum PSA concentrations.

So what about men over 50 years, with higher PSA values? - apparently no evidence on that one, yet that is exactly the patient group where evidence is most needed. It is interesting to relate this to the article on diagnostic tests in this edition of *Bandolier*.

So without wanting to preempt considerable primary study and overview work presently being undertaken in this area, it is easy to see why one possible screening programme should have the status of "not proven", despite the considerable pressures to introduce it.

References:

- 1 R Simak, S Madersbacher, Z-F Zhang, U Maier. The impact of ejaculation on serum prostate specific antigen. *Journal of Urology* 1993 150: 895-7.
- 2 WJ Glenski, GG Klee, EJ Bergstralh, JE Oesterling. PSA: establishment of the reference range for the clinically normal prostate gland and the effect of DRE, ejaculation and time on serum concentrations. *The Prostate* 1992 21:99-110.

SWOTS' CORNER

1 MALMAP - four legs good, two legs better

In the absence of very large randomised controlled trials, meta-analysis of the results of small trials remains one of the best types of evidence. As with any type of research, good work is best; as always it is useful to know what features of any piece of work we should look for to see if it is the best we could ask for.

In meta-analysis what is emerging is a clear distinction between two types of meta-analysis - that based on meta-analysis of individual patient data (MAP) and meta-analysis of data from the literature (MAL).

MAL is the traditional type of meta-analysis where the authors find all the trials on a particular topic in the literature, and then combine the data in the form of a meta-analysis. Most of the meta-analyses in the Cochrane Database of Systematic Reviews, and, indeed, most meta-analyses ever done are of this type. These are tremendously helpful - they identify all the papers published on a particular topic (and often unpublished material) as well as providing an overview of the results. Even if we do not agree with the result or conclusions, at least we can get the papers, read them, and try to draw our own conclusions.

In meta-analysis of individual patient data (MAP) those doing the meta-analysis ask research workers for the original patient data and work with and from it to derive their meta-analytic conclusions. The benefits of doing this are elegantly and succinctly described in a JAMA leader [1]; this is worth reading by anyone who uses reviews for policy or clinical decision making. A detailed paper on meta-analysis of individual patient data [2] is worth reading by those doing or thinking of doing meta-analysis.

2 Regression To the Mean

The efficacy of three interventions to lower plasma lipid levels was explained in a randomised controlled trial. In all three groups cholesterol concentration fell by about 10% after six months and the researchers concluded that the interventions had been the cause of the reduction. The explanation could, however, have been simply that there had been regression to the mean (RTM).

The words regression to the mean can bring an overpowering weariness to those of us unused to statistical terms. The phenomenon occurs with any physiological variable; for cholesterol, for instance, adults have an average daily peak-to-trough variation of 0.7 mmol/L (range 0.4 - 0.9 mmol/L), but the daily range in some diabetic patients can be as high as 1.9 mmol/L [3].

Blood pressure varies from day to day and hour to hour. If a sample of people is drawn from the population because they have "high blood pressure" then some of them may simply be at the upper end of fluctuation and when they are measured again will have a lower blood pressure. If people were

selected for low blood pressure (a common diagnosis in continental Europe) a proportion of them will be found to have a higher pressure at later times as they regress to their mean.

RTM is at the same time both a simple and a subtle phenomenon, but is excellently explained in an article in the Lancet [4]. This paper is a bit of a mindstretcher, and some of the maths will be beyond many of us, but none of the text is too difficult. Clinicians who deal with individual patients and/or with variable like blood pressure or lipids, (and those who do not), should read this paper and add it to their own swots' corner collection.

References:

- 1 AD Oxman, MJ Clarke, LA Stewart. Meta-analyses using individual patient data are needed. *Journal of the American Medical Association* 1995 274: 845-6.
- 2 LA Stewart, MJ Clarke. Practical methodology of meta-analyses (overviews) using updated individual patient data. *Statistics in Medicine* 1995 14:2057-79.
- 3 RW Simpson, RD Carter, RA Moore, WAF Penfold. Diurnal changes in plasma lipoproteins in normal subjects and diabetics. *Diabetologia* 1980 18:35-40.
- 4 P Yudkin, IM Stratton. How to deal with regression to the mean in intervention studies. *Lancet* 1996 347: 241-3.